Matrix balancing based interior point methods for point set matching problems^{*}

Janith Wijesinghe^{$\dagger 1$} and Pengwen Chen^{$\ddagger 2$}

^{1,2}Applied Mathematics, National Chung Hsing University, Taichung, 402, Taiwan.

February 17, 2023

Abstract

Point sets matching problems can be handled by optimal transport. The mechanism behind it is that optimal transport recovers the point-to-point correspondence associated with the least curl deformation. Optimal transport is a special form of linear programming with dense constraints. Linear programming can be handled by interior point methods, provided that the involved ill-conditioned Hessians can be computed accurately. During the decade, matrix balancing has been employed to compute optimal transport under entropy regularization approaches. The solution quality relies on two factors: the accuracy of matrix balancing and the boundedness of the dual vector. High accurate matrix balancing is achieved by the application of Newton methods on a sequence of matrices along a central path. In this work, we apply sparse support constraints to matrix-balancing based interior point methods, in which the sparse set fulfilling total support is iteratively updated to truncate the domain of the transport plan. Total support condition is one crucial condition, which guarantees the existence of matrix balancing as well as the boundedness of the dual vector.

Keywords: Optimal transport, interior point methods, matrix balancing, negative entropy, point-set matching problems

1 Introduction

Registration aims to match two or more sets of image data altered by geometric transforms, taken at different times, or from different sensors. Point set representing image data is commonly employed to reduce the computational load in computer vision. The associated point-set matching problem (registration) is to establish a consistent point-to-point correspondence between two point sets and to estimate the spatial alignment transformation. The quality of correspondence plays a crucial role in estimating followup transformations in registration. The iterative closest point (ICP) algorithm is one classic and popular approach in feature-based image registration problems, because of its simplicity [BM92]. For correspondence correctness, the ICP algorithm requires sufficient overlap between the point sets. Its vulnerability in performance also includes the proneness to outliers. To alleviates these difficulties, researchers describe the correspondence by a permutation matrix, which minimizes some "distance" of the point-sets, typically consisting of one regularization term for transformations and one assignment term for correspondence. For instance, Chui and Rangarjan proposed a robust point matching method(RPM), which estimates non-rigid transformation and correspondence simultaneously, where the point-to-point correspondence is enforced by Sinkhorn matrix balancing [CR00]. This can be viewed as one early application of optimal transport in registration. Comprehensive surveys of traditional registration methods can be found in [MV98] and [ZF03].

Two unlabeled point sets can be regarded as two histograms, whose distance can be fast evaluated by various information divergences, e.g., Hellinger distances, Kullback-Leibler divergences and Jensen

 $^{^{*}}$ Submitted to the editors DATE. Funding: The research of the author is supported by grant 110-2115-M-005-007- MY3 from the Ministry of Science and Technology, Taiwan.

[†]Email address: janithuop@gmail.com

[‡]Corresponding author. Email address: pengwen@nchu.edu.tw, pengwen@email.nchu.edu.tw

Shannon divergences. From a perspective of correspondence retrieval, a natural choice is Wasserstein distance (also known as the earth mover's distance [RTG00b]). Wasserstein distance quantifies the minimal cost of moving the probability mass from one distribution to the other distribution. In the 1780's, Monge described a problem of transporting a pile of soil with the least amount of work. In the 1940's, Kantorovich [Kan42] employed a dual variation principle to convert the original nonlinear problem into a linear programming problem and to study the optimal solutions. A survey of theoretical works on this problem can be found in [Eva97] [Vil03] or [Vil08]. Nowadays, optimal transport has been applied in various tasks, including image retrieval, image registration, image morphing, shape matching and maching learning, see [WPR85], [Kai98], [RTG00a], [ZYHT07], [RDG09], [PC19], [SS15], [MSKL09], [CLC13], [KPT⁺17], [CA14].

In the application of point-set registration, we can incorporate optimal transport in feature based methods to estimate the transforms and the correspondence in the existence of outliers, for instance, Hellinger distances based point set matching model (HD) [Che11a]. The HD model can be regarded as an approximation of optimal transport, when the kernel scale tends to infinity. With a finite kernel scale, the measure preserving constraint is relaxed to tolerate the existence of outliers. The effectiveness of this application generally depends on the hypothesis of geometric transforms. A fundamental question is, for which class of transformations the underlying point correspondence can be reconstructed correctly? Impressively, when the transformation can be expressed as the gradient of some convex function, the underlying correspondence can be recovered correctly by solving the L^2 optimal transport problem. The set of transformations includes scalings, translations, positive definite affine transforms and other curlfree maps. This property makes optimal transport models suitable and robust in certain applications. For instance, [CLC13] applies the optimal mass transport model to match lung vessel branch points, which are extracted from two computed tomography(CT) lung images acquired during breath-holds. Although the physical deformation field is rather large and complex, the correspondence reconstruction is surprisingly almost perfect, which verifies the superiority of the optimal transport model.

Despite of the theoretical advantage, optimal transport is limited by its heavy computational requirement in practical applications. Briefly, as one member of linear programming, optimal transport can be solved by various algorithms in linear programming. Standard algorithms include the simplex method and the interior point method [Rob12] [LY16] [Gon12]. Nowadays the primal-dual interior method is an efficient interior point method in solving linear programming, when the problem size is moderate [Wri97]. Thanks to second-order convergence in each sub-problem, an interior point method can quickly generate accurate solutions from proper matrix-free algorithms. For instance, in the community of machine learning, Wasserstein barycenter is one average of multiple discrete probability measures in terms of Wasserstein distance [YLST21, GWXY19], where accurate solutions can be computed by interior point methods [GWXY19]. In considering the flexibility of handling transformation and correspondence simultaneously, we focus on the negative entropy function as a regularizer to handle optimal transport in the registration problem. This regularization elegantly converts optimal transport to one matrix balancing task. Actually, matrix balancing algorithms are known as an effective tool to produce one approximation of the optimal transport plan [Cut13] [BCC⁺15] [KR17] [Sch19]. The major numerical tool is the Sinkhorn balancing algorithm [Sin64] [KS67]. To improve the convergence speed of Sinkhorn algorithm, the ϵ -scaling heuristic and the kernel truncation are introduced to reduce the number of iterations and the number of variables to reduce the computational load [Sch19].

1.1 Contributions

This paper is concerned with the application of this matrix balancing based interior point methods in solving point-set matching problems. The main question is whether we can develop a proper central path for discrete optimal transport approximations with small regularization parameters? The contribution can be summarized as follows. First, we investigate the application of Newton methods in matrix balancing based interior point methods for optimal transport. Although Sinkhorn balancing algorithm is popular and widely used in balancing matrices, it is generally difficult to produce an accurate result quickly for our application. In this paper, we propose Sinkhorn-Newton Negative entropy interior point methods. One underlying challenging is that as the central path heads toward an optimal permutation solution, the rank of the associated Schur complement matrices reduces to n, where n is the point cardinality in each point-set. During the rank-reduction process, it is numerically challenging to maintain the

accuracy of Newton iterates. To overcome this, we adopt the techniques proposed in the stabilized scaling algorithm [Sch19], including computations in the Log-domain and the translation of scaling vectors. See section 3.4.

Second, we revisit a few matrix balancing algorithms, including the Knight-Ruiz(KR) fixed point method [KR12]. Our matrix-balancing experiments confirm the excellent performance of KR algorithm, although its global convergence is unclear. To reveal the connection between KR and other Newton methods, we introduce one convex function for matrix balancing task and propose a novel modified Newton method, called LB algorithm. The KR algorithm is the modified Newton method with step size 1. Theorem 3 indicates that when LB is applied to a matrix with total support, the step size will be 1, as the iterates get close to an optimal solution.

Third, as in the kernel truncation method [Sch19], sparse support sets can be imposed to reduce the memory requirement in the application of the interior point methods to large-scale problems. However, the truncated kernel matrix does not always have total support, which is crucial to guarantee the quality of matrix balancing computation and the boundedness of the scaling vectors. In Prop. 2.3, we propose one simple method to construct one sparse support set with total support, and propose SNNE-sparse in Alg. 2.5, which are cable of handling large-scale matching problems. To evaluate sparse support matrix balancing methods, Theorem 2 gives one error bound estimate, which relates the boundedness of the dual vector to the duality measure estimate. According to Remark 3.8, the boundedness of the dual vector can be ensured, if the truncated matrix satisfies the total support condition.

This paper is organized as follows. In section 2, we describe the application of optimal transport in point-set registration. Discrete optimal transport can be solved by matrix balancing based interior point methods, including SNNE and SNNE-sparse. In section 3, we describe a few matrix balancing schemes, including Sinkhorn-Knopp balancing, Knight-Ruiz scheme and other Newton methods. Matrix balancing can be achieved through minimizing a convex function. In section 4, we present a few numerical simulations, which demonstrate the effectiveness of the proposed algorithms SNNE and SNNE-sparse.

1.2 Notations

In this paper, let $\langle x, y \rangle$ denote the inner product between x, y in \mathbb{R}^n . For a vector $x \in \mathbb{R}^n$ and a scalar $\epsilon \in \mathbb{R}$, let $y = (x > \epsilon)$ denote a zero-one vector, i.e., for $i = 1, \ldots, n$, set $y_i = 1$ if $x_i > \epsilon$, and set $y_i = 0$ otherwise. For simplicity of notation, the functions exp and log are extended to vector spaces \mathbb{R}^n by componentwise application to all components: $(\exp(x))_i = \exp(x_i), (\log x)_i = \log x_i, i = 1, \ldots, n$. Likewise, let x^{-1} be the vector whose entries are x_i^{-1} . Let the operator \odot denote entrywise multiplication, e.g., $x \odot y \in \mathbb{R}^n$ and $(x \odot y)_i = x_i y_i$. Let $\mathbb{1}_n = [1, 1, \ldots, 1]^\top \in \mathbb{R}^n$ be the vector whose entries are all one. Let [x; y] denote the stacked vector $[x^\top, y^\top]^\top$ for any two vectors x, y. The norm $\|\cdot\|$ represents the 2-norm. Let \mathbb{T} be the reshape operator $x \in \mathbb{R}^{n^2} \to \mathbb{R}^{n \times n}$, $\mathbb{T}(x) \in \mathbb{R}^{n \times n}$, $\mathbb{T}(x)_{i,j} = x_{in+j}$ for $i, j \in \{1, \ldots, n\}$. In addition, for the sake of simplicity, $x_{i,j}$ stands for $\mathbb{T}(x)_{i,j}$ if no confusion occurs. Let Π_n denote the set of doubly stochastic matrices, i.e., row stochastic and column stochastic $\mathbb{T}(x)\mathbb{1}_n = \mathbb{1}_n = \mathbb{T}(x)^\top \mathbb{1}_n$ for each $\mathbb{T}(x) \in \Pi_n$. Finally, A^{\dagger} stands for the pseudo inverse of a matrix A.

2 Optimal transport

2.1 Matching point-sets under deformations

We first review the deformation characterization of optimal transport applied on the point-set matching problems in the previous work [CLC13]. The primary focus of the point set matching is the reconstruction of the correspondence between two unlabeled point-sets $\{z_i\}_{i=1}^n \subset \Omega$ and $\{y_i\}_{i=1}^n \subset T(\Omega)$, where T is some injective and orientation-preserving deformation on a bounded open connected subset Ω of \mathbb{R}^3 . The correspondence can be described by a permutation τ such that $y_i = T(z_{\tau(i)})$ and some optimal condition hold for τ . One natural criterion is the minimization problem:

$$\min_{\tau} \sum_{i=1}^{n} \|y_i - z_{\tau(i)}\|^2.$$
(1)

This is a discrete combinatorial optimization problem, because n! possibilities must be evaluated. This difficulty can be alleviated, if we consider the relaxed continuous problem,

$$\min_{X_{i,j}} \sum_{i=1}^{n} \|y_i - z_j\|^2 X_{i,j},\tag{2}$$

subject to the unit mass constraints $\sum_{i=1}^{n} X_{i,j} = 1 = \sum_{j=1}^{n} X_{i,j}$ and $X_{i,j} \ge 0$. The problem is known as the L^2 Monge-Kantorovich mass transport problem. The relaxed problem described by Eq. (2) is a convex (in fact, linear) minimization problem, which has an optimal permutation matrix (the existence of this is guaranteed by Birkhoff's theorem) and can be solved by interior point methods [BV04] or primal-dual algorithms [Kai98] (see chapter 4 in [BDM09]).

In the context of (1), the permutation τ corresponding to the permutation X is optimal, if and only if $\{(z_{\tau(i)}, y_i)\}_{i=1}^n$ is cyclically monotone. Consider a transform $T : \mathbb{R}^d \to \mathbb{R}^d$ between two point sets $\{z_i\}_{i=1}^n, \{y_i\}_{i=1}^n$ in \mathbb{R}^d with $y_i = T(z_i)$. When a (unknown) transform between these point sets is the gradient of some convex function, then the correspondence can be recovered correctly by solving mass transport problems. The set of transforms includes scalings, translations, and other curl-free maps. Point correspondence can be reconstructed correctly from optimizing transport objectives, if the transform T between point-sets is the gradient of some convex function. In general, for a point-set with finite cardinality n sampled from $\Omega \subset \mathbb{R}^3$, when the curl of the transform is sufficiently small, then the underlying correspondence coincides with a minimizer $\{X_{i,j}\}_{i,j=1}^n$ of Eq. (2). Empirical studies show the outstanding performance of optimal transport in recovering the point-to-point correspondence under a small curl deformation [CLC13].

2.2 Discrete optimal transport

To solve (2), introduce a vector $c \in \mathbb{R}^{n^2}$ and its associated (reshaped) matrix $\mathbb{T}(c)$ with $\mathbb{T}(c)_{i,j} = ||y_i - z_j||^2$. We can express (2) as the primal problem (transportation): searching for the optimal solution $x \in \mathbb{R}^{n^2}$ in

$$\min_{\mathbb{T}(x)\in\Pi_n} \langle c, x \rangle, \tag{3}$$

where Π_n is the set

$$\{\mathbb{T}(x): Mx := [\mathbb{T}(x)\mathbb{1}_n; \mathbb{T}(x)^\top \mathbb{1}_n] = \mathbb{1}_{2n}, x \ge 0\}.$$
(4)

The matrix $X = \mathbb{T}(x)$ represents a coupling matrix $X = [X_{i,j} \ge 0 : i, j = 1, ..., n]$, whose entry $X_{i,j}$ describes the amount of mass flowing from bin *i* toward bin *j*. The problem in (3) is also known as the assignment problem with assignment matrix $\mathbb{T}(c)$. For each feasible solution *x*, at most *n* entries can reach the value 1, i.e., $\mathbb{T}(x)$ is a permutation matrix. By Birkhorff theorem, the extreme points of the set of doubly stochastic matrices are the permutation matrices.

The action of the adjoint operator M^{\top} on a vector $\nu = [\nu^{(1)}; \nu^{(2)}] \in \mathbb{R}^{2n}$ is given by

$$M^{\top}\nu = \mathbb{T}^{-1}(\nu^{(1)}\mathbb{1}_{n}^{\top} + \mathbb{1}_{n}\nu^{(2)^{\top}}).$$
(5)

Its dual problem to (3) is the maximization problem with respect to a dual variable $\nu \in \mathbb{R}^{2n}$,

$$\max_{\nu} \{\mathbb{1}_{2n}^{\top}\nu : M^{\top}\nu \le c\}.$$
(6)

The optimal condition of the primal and dual problem is characterized by the Karush-Kuhn-Tucker(KKT) conditions, i.e., the nonnegativeness of a slack vector in (6),

$$s := c - M^{\top} \nu \ge 0 \tag{7}$$

holds and $\mathbb{T}(s)_{i,j} > 0$ occurs only for those indices (i, j) with $X_{i,j} = 0$. The slackness condition actually implies zero duality gap,

$$\langle c, x \rangle - \langle \nu, \mathbb{1}_{2n} \rangle = \langle c, x \rangle - \langle M^{\top} \nu, x \rangle = \langle s, x \rangle = 0.$$
 (8)

2.3 Interior point methods

Here we quickly illustrate the application of interior point methods to (3). More details can be found in textbooks [BV04] and [LY16]. We start with log-barrier functions for a basic conceptual introduction of interior point methods, which motivates the negative entropy barrier functions in our interior point methods.

To reach one optimal solution of (3), path-following methods [FM68] solve the associated logarithmic barrier function with larger and larger values of $t \in \{t_j : 0 < t_0 < t_1 < t_2 < ...\}$,

$$\min_{\mathbb{T}(x)\in\Pi_n} \{ c^\top x - t^{-1} \langle \mathbb{1}_{n^2}, \log x \rangle \}.$$
(9)

For each $t = t_i > 0$, let $x = x^{(t)}$ be the critical point of the Lagrangian function,

$$\min_{x} \{ f(x,\nu) := c^{\top} x - t^{-1} \langle \mathbb{1}_{n^2}, \log x \rangle - \nu^{\top} (Mx - \mathbb{1}_{2n}) \}.$$
(10)

We compute the central point $x^{(t_j)}$ starting from the previously computed central point $x^{(t_{j-1})}$. The following proposition shows the KKT condition of (9). The proof can be given by the direct calculus.

Proposition 2.1. Consider (9) with t > 0. Introducing a multiplier vector ν for the constraint $Mx = \mathbb{1}_{2n}$, we have the Lagrangian function

$$c^{\top}x - t^{-1} \langle \mathbb{1}_{n^2}, \log x \rangle - \nu^{\top} (Mx - \mathbb{1}_{2n}).$$
 (11)

The optimal condition of x is

$$c \odot x - t^{-1} \mathbb{1}_{n^2} = \operatorname{diag}(x) M^{\top} \nu, \ i.e., tx = (c - M^{\top} \nu)^{-1},$$
 (12)

where thanks to the constraint $Mx = \mathbb{1}_{2n}$, ν is a root of the nonlinear equation,

$$M(c - M^{\top}\nu)^{-1} - t\mathbb{1}_{2n} = 0, \text{ subject to } M^{\top}\nu < c.$$
(13)

The condition in (12) states that $c \odot x - t^{-1} \mathbb{1}_{n^2}$ lies in the range of diag(x)M for the optimal interior point x > 0 in Π_n . Taking the product (12) with x yields the duality gap $t^{-1}n^2$ associated with finite t, i.e.,

$$c^{\top}x - \mathbb{1}_{2n}^{\top}\nu = t^{-1}n^2 \ge 0, \tag{14}$$

which provides a measure of closeness to optimality. The optimal solution of (3) can be obtained from a limit of $x^{(t)}$ as $t \to \infty$.

2.3.1 Matrix-free conjugate gradient methods for central path

We illustrate the matrix-free computation of $x^{(t)}$. The argument is standard, for instance, see [LY16]. We start with one initial point $x^{(t_0)}$ in Π_n . To approximate the critical point $x^{(t)}$ in (10), we generate a minimizing sequence $\{(x_k, \nu_k) : k = 1, 2, 3, ...\}$ of (10) with step size $\alpha > 0$,

$$x_{k+1} = x_k + \alpha d_k \in \mathbb{T}^{-1}(\Pi_n), \ \nu_{k+1} = \nu_k + \alpha y_k,$$
(15)

where $z_k := (d_k, y_k)$ satisfies the linearization of (10)

$$\nabla f(x_k + d_k, \nu_k + y_k) \approx \nabla f(x_k, \nu_k) + \langle \nabla f^2(x_k, \nu_k), z_k \rangle = 0.$$
(16)

Introduce the residual vector,

$$r_k = -(c - (x_k t)^{-1} - M^{\top} \nu_k).$$
(17)

Together with $Mx_k = \mathbb{1}_{2n}$, (16) gives

$$\nabla^2 f(x_k,\nu_k)z_k = \begin{pmatrix} t^{-1}\operatorname{diag}(x_k)^{-2}, & -M^\top \\ -M, & 0 \end{pmatrix} \begin{pmatrix} d_k \\ y_k \end{pmatrix} = -\nabla f(x_k,\nu_k) = \begin{pmatrix} r_k \\ 0 \end{pmatrix}.$$
(18)

The first part of (18) implies

$$t^{-1}d_k = \text{diag}(x_k^2)(M^{\top}y_k + r_k).$$
(19)

Together with the second part of (18), we have the normal equation for y_k ,

$$M \operatorname{diag}(x_k^2)(M^{\top} y_k + r_k) = 0, \ i.e., y_k = -(M \operatorname{diag}(x_k^2)M^{\top})^{\dagger}(M \operatorname{diag}(x_k^2)r_k).$$
(20)

The well-poshness of (20) is given in the appendix. We can employ Krylov subspace methods, e.g., matrix-free conjugate gradient methods to solve y_k from (20) and then compute d_k from (19).

In solving (18), we shall avoid forming those big matrices M and $\operatorname{diag}(x_k^{-2})$. We demonstrate the matrix-vector product in the conjugate gradient method in solving y_k . With $y_k := [y^{(1)}; y^{(2)}]$ and

$$\widetilde{M}_k := M \operatorname{diag}(x^2) M^\top = \begin{pmatrix} \operatorname{diag}(\mathbb{T}(x_k^2) \mathbb{1}_n), & \mathbb{T}(x_k^2) \\ \mathbb{T}(x_k^2)^\top, & \operatorname{diag}(\mathbb{T}(x_k^2)^\top \mathbb{1}_n) \end{pmatrix},$$
(21)

we implement the matrix-vector product in the conjugate gradient method,

$$\widetilde{M}_k y_k = \begin{pmatrix} y^{(1)} \odot (\mathbb{T}(x_k^2) \mathbb{1}_n) + \mathbb{T}(x_k^2) y^{(2)} \\ y^{(2)} \odot (\mathbb{T}(x_k^2)^\top \mathbb{1}_n) + \mathbb{T}(x_k^2)^\top y^{(1)} \end{pmatrix}.$$
(22)

To further enhance the convergence speed, we can adopt some preconditioners for the conjugate gradient method, e.g., modified Cholesky preconditioners [FO08].

Remark 2.2 (Rank reduction). Note that the matrix $M_k := M \operatorname{diag}(x_k^2) M^{\top}$ can be regarded as the **Schur** complement of the first block in the Hessian matrix in (18), after ignoring the scaling factor t. (This matrix also appears in the Hessian computation in (67) and (84) for matrix balancing algorithms.) Each $x^{(t)}$ is computed based on the Newton direction d_k , whose calculation is essentially the application of a projection P_k . The calculation could be inaccurate, if the involved **Schur complement** $M \operatorname{diag}(x_k^2) M^{\top} \in$ $\mathbb{R}^{2n \times 2n}$ has serious rank deficiency due to the limitation of finite precision. Since the null space of M^{\top} has dimension 1, the rank of $M \operatorname{diag}(x_k^2) M^{\top}$ is 2n - 1 for x_k with all entries away from 0 (See the appendix). When the optimal solution $\mathbb{T}(x)$ of (3) is a permutation matrix, $M \operatorname{diag}(x^2) M^{\top}$, which is the sum of n rank one matrices, has rank only n. Hence, as $x^{(t)}$ tends to x, many entries in $(x^{(t)})^2$ (though nonzero) will be rounded to zero in the matrix-vector-product calculation. The inaccuracy is always inevitable for t sufficiently large. To reduce numerical errors caused by the singularity, the matrix \widetilde{M}_k should be replaced with a regularized matrix

$$M_k + \epsilon I_{2n \times 2n}$$
 for some positive small ϵ . (23)

In addition, to ensure the feasibility of x_k , we can apply matrix balancing to project x_k on Π_n . Another manner to alleviate the rank deficiency is that we can employ some early termination condition stated in Prop. A.2 to produce an optimal solution fulfilling the KKT condition, if it is applicable.

The aforementioned log-barrier interior point method only serves for the purpose of illustrating the overall algorithmic framework, and motivating the negative-entropy based interior point methods. Computational experiments show that primal-dual methods can perform much better than this pure primal barrier methods on practical problems. For instance, Mehrotra predictor-corrector method [Meh92] is one popular primal-dual method, whose iterates follow a path with duality measure tending to 0 to reach one point fulfilling the KKT condition in the space of x, ν and s [Wri97].

2.4 Optimal transport by matrix balancing

Recently, optimal transport has been approximated by an entropic regularized optimal transport problem [Cut13] [CPSV18]. Using the negative entropy function $x \log x - x$, we obtain a regularized problem with t > 0,

$$\min_{Mx=\mathbb{1}_{2n}} \left\{ \mathbb{F}_t(x) := \langle c, x \rangle + t^{-1} \langle \mathbb{1}_{n^2}, x \odot \log x - x \rangle \right\}.$$
(24)

The strict convexity of $x \odot \log x$ implies the uniqueness of the minimizer in $\mathbb{R}^{n^2}_+$. The first-order optimal condition suggests that the optimal solution can be computed by matrix scaling algorithms. Introducing

a multiplier vector ν for the constraint, we have the problem

$$\min_{\mathbf{z}} \left\{ \langle c, x \rangle + t^{-1} \langle \mathbb{1}_{n^2}, x \odot \log x - x \rangle - \langle \nu, Mx - \mathbb{1}_{2n} \rangle \right\}.$$
(25)

The gradient computation gives the optimal condition of x,

$$c + t^{-1}\log x - M^{\top}\nu = 0$$
, i.e., $x = \exp(-t(c - M^{\top}\nu)).$ (26)

The multiplier vector $\nu := [\nu^{(1)}; \nu^{(2)}]$ in (25) can be determined in matrix balancing of $\exp(-t\mathbb{T}(c))$. Indeed, since $\mathbb{T}(M^{\top}\nu) = \nu^{(1)}\mathbb{1}_n^{\top} + \mathbb{1}_n\nu^{(2)^{\top}}$, then (26) yields that $\mathbb{T}(x) \in \Pi_n$ is obtained under proper scaling matrices,

$$\mathbb{T}(x) = \mathbb{T}(\exp(-t(c - M^{\top}\nu))) = \operatorname{diag}(\exp(t\nu^{(1)}))\exp(-t\mathbb{T}(c))\operatorname{diag}(\exp(t\nu^{(2)})) \in \Pi_n.$$
(27)

Various Newton methods can be employed to perform matrix balancing in (27). The details of matrix balancing algorithms will be presented in next section.

Under large t, the solution in (24) can provide a better approximation to the original optimal transport in (3). However, problems with large t are generally very ill-conditioned and hard to solve. To alleviate the ill-condition issue, with $\eta > 1$, we solve ν in a sequence of subproblems associated with $t = t^{(0)}, t^{(1)}, \ldots, t_{\text{max}}$. This method is known as ϵ -scaling heuristic [Sch19] with 1/t replaced with $\epsilon \to 0$. To emphasize the usage of Newton methods, we call the interior point method in solving (24) with $t \to \infty$ as the **Sinkhorn-Newton-negative-entropy method(SNNE)**.

- Initialize $t = t_0$ and $\nu = \nu_{ini}$. Repeat the following two steps until $t = t_{max}$.
- Employ Newton based matrix balancing algorithms to update ν , i.e., $\exp(-t\mathbb{T}(c-M\nu))$ is doubly stochastic.
- If $t < t_{\text{max}}$, update $t \to t\eta$.

The convergence of SNNE consists of two parts: the duality gap and the slackness condition. The convergence of duality gap requires the boundedness of ν , which is related to the total support condition. We postpone the discussion to Theorem 2. Here, we give a few words on the convergence of $s \odot x \to 0$ as $t \to \infty$. With $s = c - M^{\top}\nu$, the optimal condition in (26) can be expressed as $s = -t^{-1}\log x \ge 0$. Fixing $\gamma' \in (0, 1)$ and $\gamma'' \in (1, \infty)$, we can compute an approximate solution x with

$$\gamma' t^{-1}(-x \odot \log x) \le s \odot x \le \gamma'' t^{-1}(-x \odot \log x) \tag{28}$$

As $t \to \infty$, we reach the KKT condition in (7),

$$0 \le x \odot s = -t^{-1}x \odot \log x \le (et)^{-1} \to 0.$$
(29)

As t gets sufficiently large, a solution satisfying the slackness condition can be reached with the aid of early termination in Prop. A.2. Empirically, the convergence for large t does require fast convergence and high accuracy of matrix balancing algorithms.

2.4.1 Interior point methods with total support constraints

Although an optimal solution x could be sparse, interior point methods require memory storage $O(n^2)$ for x, which could be prohibited in large-scale point-sets. As column generation solves large linear programming, we shall use dual variables to reduce the memory storage by imposing (and dynamically updating) the **sparse support constraint** on x. For instance, in [Sch19] sparse support sets are introduced to form approximate problems with truncated sparse kernels to reduce the memory storage requirement. Actually, introducing these constraints to remove those inactive components can also improve the quality of solutions $x^{(t)}$.

Let supp(x) be the index set of all the positive entries in x. We say that the index set Σ is one support of $X = \mathbb{T}(x)$, if Σ consists of all indices of nonzero entries in X, i.e., $X_{i,j} = 0$ holds for all $(i, j) \notin \Sigma$. We say that X is a solution to optimal transport with respect to the support constraint Σ , if Σ is a support of $X = \mathbb{T}(x)$ and x is one optimal solution to

$$\min_{Mx=\mathbb{1}_{2n}} \left\{ \mathbb{F}_t(x,\Sigma) := \langle c,x \rangle + t^{-1} \langle \mathbb{1}_{n^2}, x \odot \log x - x \rangle \right\}, \ supp(\mathbb{T}(x)) \subset \Sigma \}.$$
(30)

To reach one optimal transport approximation, we shall generate a sequence of supports

$$\{\Sigma_1, \dots, \Sigma_{\xi}, \dots\},\tag{31}$$

and apply matrix balancing algorithms to get an approximate solution $X_{\xi} \in \Pi_n$ with respect to the support Σ_{ξ} for each $\xi \in \{1, 2, ...\}$. By updating x and Σ alternately, we can reach a good approximation of the optimal solution for $\mathbb{F}_t(x)$ in (24). if the selection rule of $\Sigma_{\xi+1}$ is given by (35) to fulfill two conditions: the total support condition (see Definition 1) and the inclusion of the index set

$$\Sigma'' := \{(i,j) : s_{i,j} := c_{i,j} - (\nu^{(1)}(i) + \nu^{(2)}(j)) \le \epsilon\} \subset \Sigma_{\xi+1}.$$
(32)

Here, ϵ is some positive parameter to ensure the sparsity of the support.

2.4.2Total support condition

Definition 1. Let X be an $n \times n$ matrix and σ be a permutation of $\{1, 2, \ldots, n\}$. Then the sequence $\{X_{1,\sigma(1)}, X_{2,\sigma(2)}, \ldots, X_{n,\sigma(n)}\}$ is a diagonal of X (corresponding to σ). Then a nonnegative square matrix X is said to have support if X contains one positive diagonal. Also, X has total support if $X \neq 0$ and if every positive entry of X lies on a positive diagonal [KS67]. Let $\mathbb{1}_{\Sigma}$ denote the indicator matrix, whose (i, j)-entry is 1 for each $(i, j) \in \Sigma$. We say that an index set Σ satisfies total support condition, if the associated indicator matrix $\mathbb{1}_{\Sigma}$ has total support.

When $\mathbb{1}_{\Sigma}$ has no support, then $\mathbb{1}_{\Sigma}$ can not be scaled to a doubly stochastic matrix. Actually, by Birkhorff theorem, any doubly stochastic matrix is convex combination of permutation matrices. Since the support of one nonnegative matrix remains invariant under the product of positive diagonal matrices, having total support is one necessary condition for matrix balancing. Indeed, Theorem 1 states that total support is the crucial condition to ensure the existence of a doubly stochastic matrix from a sparse nonnegative matrix X.

Theorem 1. [KS67] Let X be a nonnegative squared matrix. A necessary and sufficient condition that $B = \operatorname{diag}(y) X \operatorname{diag}(z)$ is double stochastic for two positive vectors y, z is that X has total support.

To illustrate the importance of total support, consider the following example. Let $X = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 3 & 0 \\ 0 & 0 & 4 \end{pmatrix}$.

Since the entry 2 is not contained in a positive diagonal, X cannot be scaled to a doubly stochastic matrix.

However, when $X = \begin{pmatrix} 1 & .05 & 0 \\ 2 & 3 & 0 \\ 0 & 0 & 4 \end{pmatrix}$, the entry 2 is contained in the positive diagonal [0.05, 2, 4] and

thus the matrix X can be balanced. On the other hand, let $X = \begin{pmatrix} 1 & \epsilon \\ 1 & 1 \end{pmatrix}$ with $\epsilon > 0$. Even though X can be scaled to a doubly stochastic metric. can be scaled to a doubly stochastic matrix,

diag([1,t])
$$\begin{pmatrix} 1 & \epsilon \\ 1 & 1 \end{pmatrix}$$
 diag((1+t)^{-1}[1,t^{-1}]) = $\frac{1}{1+t} \begin{pmatrix} 1 & \epsilon t^{-1} \\ t & 1 \end{pmatrix}$ with $t = \epsilon^{1/2}$, (33)

the relative magnitude of entries of the scaling vectors tend to ∞ as $\epsilon \to 0$.

We illustrate the construction of a set with total support. Let Σ'' denote the index set,

$$\Sigma'' := \{ (i,j) : c_{i,j} - (\nu^{(1)}(i) + \nu^{(2)}(j)) \le \epsilon \}.$$
(34)

In general, the set Σ'' does not automatically meet the total support condition. Here is one simple construction of a total support set Σ containing the prescribed index set Σ'' .

Proposition 2.3. Let Σ'' be some prescribed index set. Let σ be a permutation of $\{1, 2, ..., n\}$ and let $\Sigma' := \{(i, \sigma(i)) : i = 1, ..., n\}$. Then the union set

$$\Sigma := \Sigma' \cup \Sigma'' \cup \Sigma''', \ \Sigma''' := \{ (\sigma^{-1}(j), \sigma(i)) : (i, j) \in \Sigma'' \}$$

$$(35)$$

has total support.

Proof. For each $(i, j) \in \Sigma''$, we shall point out one diagonal in Σ . Since σ is a permutation, then $\{(k, \sigma(k)) : k = 1, 2, 3, ..., n\}$ is one diagonal. Express the diagonal sequence as $\{(i, \sigma(i)), (\sigma^{-1}(j), j), \widehat{\Sigma}\}$, i.e., $\widehat{\Sigma}$ is the set consisting the remaining n-2 indices. Note that $\widehat{\Sigma}$ does not consist of any entries in row-i, row- $\sigma^{-1}(j)$, column-j and column- $\sigma(i)$. Then $\{(i, j), (\sigma^{-1}(j), \sigma(i)), \widehat{\Sigma}\}$ is a diagonal for this (i, j). \Box

Remark 2.4 (The choice of σ). The set Σ''' can be regarded as one "reflection" of Σ'' with respect to the diagonal Σ' . For simplicity, one can consider the fixed choice: let σ to be the identity and Σ in (35) is the index set corresponding to the positive entries of $I + \mathbb{1}_{\Sigma''} + \mathbb{1}_{\Sigma''}^{\top}$. Empirically, we suggest that the permutation σ should be chosen dynamically, so that the corresponding entries $\{X_{i,\sigma(i)} : i = 1, \ldots, n\}$ are large entries in X, away from zero.

2.4.3 Index set Σ''

The inclusion of Σ'' is to provide one tight approximation to $\mathbb{F}_t(x)$ in (24). Substitute the optimal vector x in (26) to (25). The Lagrange dual of (24) is given by

$$\max_{\nu} \left\{ \mathbb{G}_t(\nu) := -t^{-1} \langle \exp(t\nu^{(1)}), \exp(-t\mathbb{T}(c)) \exp(t\nu^{(2)}) \rangle + \langle \nu, \mathbb{1}_{2n} \rangle \right\}.$$
(36)

Introduce a sparse support set Σ as the support of x and solve x from the problem

$$\min_{Mx=\mathbb{1}_{2n}, x \ge 0} \left\{ \mathbb{F}_t(x, \Sigma) := \langle c, x \rangle + t^{-1} \langle \mathbb{1}_{\Sigma}, x \odot \log x - x \rangle \right\}.$$
(37)

Introduce a multiplier vector ν for the constraint and form the Lagrangian function,

$$\langle c, x \rangle + t^{-1} \langle \mathbb{1}_{\Sigma}, x \odot \log x - x \rangle - \langle \nu, Mx - \mathbb{1}_{2n} \rangle.$$
(38)

The optimal solution is given by

$$x = \mathbb{T}^{-1}(\mathbb{1}_{\Sigma}) \odot \exp(-t(c - M^{\top}\nu)), \quad \text{i.e.,} \quad \mathbb{T}(x) = \operatorname{diag}(\nu^{(1)})(\mathbb{1}_{\Sigma} \odot \exp(-t\mathbb{T}(c)))\operatorname{diag}(\nu^{(2)}), \tag{39}$$

where ν is chosen to ensure $\mathbb{T}(x) \in \Pi_n$. Using (39), we have the Lagrange dual of (37),

$$\max_{\nu} \left\{ \mathbb{G}_t(\nu, \Sigma) := -t^{-1} \langle \mathbb{1}_{\Sigma}, \exp(-t\mathbb{T}(c - M^{\top}\nu)) \rangle + \langle \nu, \mathbb{1}_{2n} \rangle \right\}.$$
(40)

Let Π_0 denote the whole index set $\{(i, j) : 1 \leq i, j \leq n\}$ and let Σ^c denote the complement set of Σ . According to duality,

$$\max_{\nu} \mathbb{G}_t(x, \Pi_0) = \min_{x \in \Pi_n} \mathbb{F}_t(x) = \min_{x \in \Pi_n} \mathbb{F}_t(x, \Pi_0) \le \min_{x \in \Pi_n} \mathbb{F}_t(x, \Sigma) = \max_{\nu} \mathbb{G}_t(\nu, \Sigma).$$
(41)

Hence, $\max_{\nu} \mathbb{G}_t(\nu, \Sigma)$ is one upper estimate for $\min_x \mathbb{F}_t(x)$ and the gap can be estimated by

$$\max_{\nu} \mathbb{G}_t(x, \Sigma) - \max_{\nu} \mathbb{G}_t(x, \Pi_0) \le \max_{\nu} (\mathbb{G}_t(x, \Sigma) - \mathbb{G}_t(x, \Pi_0)) \le \max_{\nu} \{ t^{-1} \langle \mathbb{1}_{\Sigma^c}, \exp(-t\mathbb{T}(c - M^\top \nu)) \rangle \}.$$
(42)

For a tight estimate to $\min_{x \in \Pi_0} \mathbb{F}_t(x)$, the support set Σ should be chosen to include the index set $\{(i, j) : (c - M^\top \nu)_{i,j} < \epsilon\}$ for some constant $\epsilon > 0$.

In summary, we have the following SNNE-sparse algorithm. As pointed in Theorem 1, the support set must satisfy total support condition to ensure the existence of scaling vectors $\nu^{(1)}$ and $\nu^{(2)}$ for matrix balancing.

Algorithm 2.5 (SNNE with sparse support). Input: parameters $\epsilon > 0$, $\xi_{\text{max}} > 0$, $t_{\text{max}} > 0$, $\eta > 1$ and the assignment matrix c. Initialize $t = t_0$ and ν . Generate one initial support set Σ_1 fulfilling the total

support condition. Repeat the following steps for $t = t_0, t_1, \ldots, t_{\max}$, so that ν_{ξ} gives a solution for x in (39).

- For $\xi = 1, 2, 3, \dots, \xi_{\text{max}}$, iterate the following two steps to get approximation solutions for ν, Σ .
 - 1. Employ Newton method based matrix balancing algorithms in section 3.2 to update ν , i.e.,

$$\mathbb{1}_{\Sigma} \odot \exp(-t\mathbb{T}(c - M\nu)) \tag{43}$$

is doubly stochastic.

- 2. Let $\nu_{\xi} = [\nu^{(1)}; \nu^{(2)}]$ and construct Σ'' by (34). Let $\Sigma_{\xi+1}$ be the total support set in (35).
- If $t < t_{\max}$, update $t \to t\eta$.

Remark 2.6 (Convergence). We give a few comments on the convergence of SNNE-sparse. Suppose we fix the cardinality $|\Sigma_{\xi}|$ for each ξ . The sequence of (x_{ξ}, Σ_{ξ}) is actually constructed to minimize $\mathbb{F}_t(x, \Sigma)$ alternately, where x_{ξ} is given by (39). Since the function $\mathbb{F}_t(x, \Sigma)$ is bounded below, the sequence will eventually stop at some ξ . Indeed, the optimality of x_{ξ} is ensured if $\mathbb{T}(x_{\xi})$ in (39) is balanced by some ν_{ξ} . From (37), the optimality of Σ_{ξ} is ensured, if Σ_{ξ} contains the index set associated with the smallest entries of $x \log x - x$, equivalently, the smallest entries of $c - M^{\top}\nu$. (Thanks to the monotonic decrease of $x \log x - x$ for $x \in [0, 1]$, Σ_{ξ} actually contains the index set associated with the largest entries of x.) Here, we ignore the total support requirement on each Σ_{ξ} .

2.4.4 Error estimate of SNNE-sparse

Error estimates of SNNE-sparse can be examined by duality measure $\langle c, x \rangle - \langle \nu, \mathbb{1}_{2n} \rangle$. The following result indicates how the duality measure under $t \to \infty$ can be improved by the accuracy of matrix balancing on $\mathbb{T}(x)$ and the boundedness assumption on ν .

Theorem 2. Consider an approximate optimal solution x of (30), constructed from matrix balancing

$$x = \mathbb{T}^{-1}(\mathbb{1}_{\Sigma}) \odot \exp(-t(c - M^{\top}\nu))$$
(44)

for some dual vector ν . Let \mathcal{N} be the null space of diag $(\mathbb{T}^{-1}(\mathbb{1}_{\Sigma}))M^{\top}$ and let P be the projection with kernel \mathcal{N} . Suppose that $\|P\nu\|_2 \leq \delta$ holds for some positive constant $\delta > 0$ and $\mathbb{T}(x)$ is nearly doubly stochastic, i.e., $\|Mx - \mathbb{1}_{2n}\|_2 \leq \epsilon_{MB}$ for some $\epsilon_{MB} > 0$. Then we have error estimates,

$$|\langle c, x \rangle - \langle \mathbb{1}_{2n}, \nu \rangle| \le \epsilon \delta + (et)^{-1} |\Sigma|, \tag{45}$$

where $|\Sigma|$ is the cardinality of the index set Σ .

Proof. Since Σ has total support, then $\mathbb{1}_{\Sigma}$ can be balanced by some scaling vectors $\zeta^{(1)}, \zeta^{(2)} \in \mathbb{R}^n$, i.e., $\mathbb{1}_{\Sigma} \odot (\zeta^{(1)} \zeta^{(2)^{\top}})$ is doubly stochastic, $M \mathbb{T}^{-1}(\mathbb{1}_{\Sigma} \odot (\zeta^{(1)} \zeta^{(2)^{\top}})) = \mathbb{1}_{2n}$. Hence,

$$M(\mathbb{T}^{-1}(\mathbb{1}_{\Sigma}) \odot x) - \mathbb{1}_{2n} = M(\mathbb{T}^{-1}(\mathbb{1}_{\Sigma}) \odot (x - \mathbb{T}^{-1}(\zeta^{(1)}\zeta^{(2)^{\top}})))$$
(46)

lies in the range of $M \operatorname{diag}(\mathbb{T}^{-1}(\mathbb{1}_{\Sigma}))$, and also lies in the range of P, which implies $P(Mx - \mathbb{1}_{2n}) = Mx - \mathbb{1}_{2n}$ from the definition of P. Computation shows

$$|c^{\top}x - \nu^{\top}\mathbb{1}_{2n}| = |c^{\top}x - \nu^{\top}Mx + \nu^{\top}(Mx - \mathbb{1}_{2n})|$$
(47)

$$\leq |(c - M^{\top}\nu)^{\top}x| + ||P\nu||_{2}||Mx - \mathbb{1}_{2n}||_{2}$$
(48)

$$= -t^{-1} \langle \mathbb{T}^{-1}(\mathbb{1}_{\Sigma}), x \odot \log x \rangle + \|P\nu\|_2 \|Mx - \mathbb{1}_{2n}\|_2$$
(49)

$$\leq (et)^{-1}|\Sigma| + \delta\epsilon, \tag{50}$$

where the last inequality is derived from $x \log x \ge -e^{-1}$.

This result is consistent with empirical studies, where solving a negative entropy regularized optimal transport could be a challenging problem, if the norm of the associated dual vector is large. Later, we shall prove that the required norm bound can be obtained under the total support condition. See Prop. 3.7 and Remark 3.8.

Remark 2.7 (Parameters in SNNE-sparse). It could be not easy to choose a proper parameter $\epsilon > 0$ to meet the desired sparsity. One practicable manner is to select a parameter k > 0 and let Σ consist of those (i, j) corresponding to (at most) k smallest entries $(c - M^{\top}\nu)_{i,j}$ for each row and each column. In this manner, Σ_{ξ} consists of at most (2k + 1)n entries. In section 4.2, we shall present numerical experiments under a proper value k to demonstrate the effectiveness.

3 Matrix balancing

Let A denote a positive matrix in $\mathbb{R}^{n \times n}$. Matrix balancing [Sin64] aims to find a pair of positive scaling vectors $\{\zeta^{(1)}, \zeta^{(2)}\}$, so that the matrix balancing projection

$$A' := \operatorname{diag}(\zeta^{(1)}) A \operatorname{diag}(\zeta^{(2)})$$

is doubly stochastic, i.e.,

$$A'\mathbb{1}_n = \operatorname{diag}(\zeta^{(1)})A\zeta^{(2)} = \operatorname{diag}(\zeta^{(1)})A\operatorname{diag}(\zeta^{(2)})\mathbb{1}_n = \mathbb{1}_n,\tag{51}$$

$$A^{\prime \top} \mathbb{1}_n = \operatorname{diag}(\zeta^{(2)}) A^{\top} \zeta^{(1)} = (\operatorname{diag}(\zeta^{(1)}) A \operatorname{diag}(\zeta^{(2)}))^{\top} \mathbb{1}_n = \mathbb{1}_n.$$
(52)

The existence of $\{\zeta^{(1)}, \zeta^{(2)}\}\$ is proved in [Sin64], [KS67] for any positive matrix and any nonnegative matrix with total support, respectively. Matrix scaling methods and its various applications in scientific computing, statistics and engineering can be found in the extensive survey [Ide16] and the references therein. In general, the prescribed row sums and column sums do not have to be restricted to $\mathbb{1}_n$. See [KLRS08] and [AZLOW17]. In the section, we shall list a few matrix scaling algorithms and their variants.

3.1 Sinkhorn-Knopp balancing(SK) and Knight-Ruiz(KR) method

For the conditions in (51,52), the Sinkhorn-Knopp balancing(SK) (also known as the RAS or biproportional problem [Bac70]) is one well-known method to carry out matrix balancing on A, consisting of iterates $\{(\zeta_k^{(1)}, \zeta_k^{(2)}) : k = 1, 2, 3, ...\},\$

$$\zeta_{k+1}^{(2)} = (A^{\top} \zeta_k^{(1)})^{-1}, \ \zeta_{k+1}^{(1)} = (A \zeta_k^{(2)})^{-1}.$$
(53)

We can express (53) in a symmetric manner [Kni08]. Form one symmetric matrix \widetilde{A} from A,

$$\widetilde{A} := \begin{pmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{pmatrix} = \begin{pmatrix} 0 & A \\ A^{\top} & 0 \end{pmatrix}.$$
(54)

Let $\zeta_k := [\zeta_k^{(1)}; \zeta_k^{(2)}]$ be a sequence of the scaling vectors. When

$$\zeta_1^{(2)} = (A^\top \zeta_1^{(1)})^{-1}, \tag{55}$$

the SK algorithm in (54) can be expressed in a compact form,

$$\zeta_{k+1} = (\widetilde{A}\zeta_k)^{-1},$$

whose limit $\zeta = \lim_{k \to \infty} \zeta_k$ is actually a root of

$$\mathbf{g}(\zeta) := \zeta \odot (\widetilde{A}\zeta) - \mathbb{1}_{2n} = 0.$$
(56)

Remark 3.1 $((\rho^{(1)}, \rho^{(2)})$ -balancing). In this paper, we focus on the application of point-set matching problems and thus consider the matrix balancing with $(\mathbb{1}_n, \mathbb{1}_n)$ -balancing, i.e., the row sum and the column

sum both 1. In literatures, e.g., section 3 in [Ide16], SK algorithms can be applied to reach a matrix with row sum $\rho^{(1)}$ and column sum $\rho^{(2)}$, where $(\rho^{(1)}, \rho^{(2)})$ is not necessarily restricted to $(\mathbb{1}_n, \mathbb{1}_n)$.

To solve the roots of $\mathbf{g}(\zeta) = 0$, Knight and Ruiz [KR12] proposed one Newton method,

$$\zeta_{k+1} = \zeta_k - (\operatorname{diag}(\zeta_k)\widetilde{A} + \operatorname{diag}(\widetilde{A}\zeta_k))^{\dagger}(\zeta_k \odot (\widetilde{A}\zeta_k) - \mathbb{1}_{2n})$$
(57)

$$= \zeta_k \odot \left\{ \mathbb{1}_{2n} - (B_k + \operatorname{diag}(B_k \mathbb{1}_{2n}))^{\dagger} (B_k \mathbb{1}_{2n} - \mathbb{1}_{2n}) \right\}$$
(58)

to alleviate slow convergence of SK, where $B_k := \text{diag}(\zeta_k)A\text{diag}(\zeta_k)$ is used. Compared with the SK algorithm, the Newton approach exhibits fast convergence. However, as mentioned in [KR12], the global convergence property of (57) is theoretically unclear.

3.2 Negative entropy(NE) based matrix balancing

We describe one algorithm proposed in [CMTV17, BCLW17], which implements Newton's method for matrix balancing in (27) or in (39). To simplify the notation, consider

$$A = \exp(-t\mathbb{T}(c)) \odot \mathbb{1}_{\Sigma} \in \mathbb{R}^{n \times n}, \text{ with } t = 1$$
(59)

where Σ is the support set used in SNNE-sparse. Introduce the symmetric $2n \times 2n$ -matrix \widetilde{A} as in (54). Write the scaling vector $\exp(\nu)$ of \widetilde{A} with $\nu := [\nu^{(1)}; \nu^{(2)}]$ and $\nu^{(1)} \in \mathbb{R}^n$ and $\nu^{(2)} \in \mathbb{R}^n$. Matrix balancing on A can be solved by the convex optimization (i.e., the problem in (36)),

$$\min_{\nu \in \mathbb{R}^{2n}} \left\{ \mathbf{f}(\nu) = \frac{1}{2} \langle \exp(\nu), \widetilde{A} \exp(\nu) \rangle - \langle \mathbb{1}_{2n}, \nu \rangle \right\}.$$
(60)

Indeed, reformulate (60) as follows:

$$\mathbf{f}(\nu) = \langle \exp(\nu^{(1)}), (\exp(-\mathbb{T}(c)) \odot \mathbb{1}_{\Sigma}) \exp(\nu^{(2)}) \rangle - \langle \mathbb{1}_n, \nu^{(1)} \rangle - \langle \mathbb{1}_n, \nu^{(2)} \rangle.$$
(61)

For simplicity, let $\mathbf{B}(\nu)$ denote the scaled matrix of A,

$$\mathbf{B}(\nu) := \exp(-\mathbb{T}(c - M^{\top}\nu)) \odot \mathbb{1}_{\Sigma} = \operatorname{diag}(\exp(\nu^{(1)}))(A\operatorname{diag}(\exp(\nu^{(2)})), \text{ where } \mathbb{T}(M^{\top}\nu) = \nu^{(1)}\mathbb{1}_{n}^{\top} + \mathbb{1}_{n}\nu^{(2)^{\top}}.$$
(62)

We can express ${\bf f}$ as

$$\mathbf{f}(\nu) = \langle \mathbb{1}_n, \mathbf{B}(\nu) \mathbb{1}_n \rangle - \langle \mathbb{1}_{2n}, \nu \rangle.$$
(63)

First, a scaling vector ν with $\nabla \mathbf{f}(\nu) = 0$ yields the double stochastic matrix $\mathbf{B}(\nu)$. Indeed,

$$\nabla \mathbf{f}(\nu) = M(\exp(-(\mathbb{T}(c - M^{\top}\nu))) \odot \mathbb{1}_{\Sigma})\mathbb{1}_{n} - \mathbb{1}_{2n} = \begin{pmatrix} \exp(\nu^{(1)}) \odot (A\exp(\nu^{(2)})) \\ \exp(\nu^{(2)}) \odot (A^{\top}\exp(\nu^{(1)})) \end{pmatrix} - \mathbb{1}_{2n}(64)$$
$$= \begin{pmatrix} \mathbf{B}(\nu) - I_{n} \\ (\mathbf{B}(\nu) - I_{n})^{\top} \end{pmatrix} \mathbb{1}_{n}.$$
(65)

Second, the Hessian computation verifies the convexity of f. Computation shows

$$\nabla^{2} \mathbf{f}(\nu) = \begin{pmatrix} \operatorname{diag}((\exp(-\mathbb{T}(c - M^{\top}\nu)) \odot \mathbb{1}_{\Sigma})\mathbb{1}_{n}) & \exp(-\mathbb{T}(c - M^{\top}\nu)) \odot \mathbb{1}_{\Sigma} \\ (\exp(-\mathbb{T}(c - M^{\top}\nu)) \odot \mathbb{1}_{\Sigma})^{\top} & \operatorname{diag}((\exp(-\mathbb{T}(c - M^{\top}\nu) \odot \mathbb{1}_{\Sigma}))^{\top}\mathbb{1}_{n}) \end{pmatrix}$$
(66)

$$= \begin{pmatrix} \operatorname{diag}(\mathbf{B}(\nu)\mathbb{1}_n) & \mathbf{B}(\nu) \\ \mathbf{B}(\nu)^\top & \operatorname{diag}(\mathbf{B}(\nu)^\top \mathbb{1}_n) \end{pmatrix}.$$
(67)

The following Newton's method, called Negative entropy method(NE), employs step size given by back-tracking line search to compute a minimizer of the problem in (60), i.e.,

$$\nu_{k+1} = \nu_k - \alpha (\nabla^2 \mathbf{f}(\nu_k))^{\dagger} \nabla \mathbf{f}(\nu_k).$$
(68)

Convergence arguments are standard. See section 9.5.3 [BV04]. The following shows the consistency analysis.

Proposition 3.2. Suppose the matrix A in (59) is nonnegative and has support. Then the system

$$\nabla^2 \mathbf{f}(\nu_k) w = -\nabla \mathbf{f}(\nu_k) \tag{69}$$

is consistent for some vector $w \in \mathbb{R}^{2n}$. In addition, for nonzero $\nabla \mathbf{f}(\nu_k)$, let $u = -(\nabla^2 \mathbf{f}(\nu_k))^{\dagger} \nabla \mathbf{f}(\nu_k)$. Then we have the squared **Newton decrement**

$$\langle u, \nabla^2 \mathbf{f}(\nu_k) u \rangle = \langle \nabla \mathbf{f}(\nu_k), (\nabla^2 \mathbf{f}(\nu_k))^{\dagger} \nabla \mathbf{f}(\nu_k) \rangle > 0.$$
(70)

Proof. For each vector $w = [w^{(1)}; w^{(2)}] \in \mathbb{R}^{2n}$ with $w^{(1)} \in \mathbb{R}^n$, $w^{(2)} \in \mathbb{R}^n$, the Hessian $\nabla^2 \mathbf{f}(\nu)$ is symmetric diagonally dominant [CMTV17, AZLOW17], thus the convexity of \mathbf{f} is verified from

$$\langle w, \nabla^2 \mathbf{f}(\nu)w \rangle = \sum_{i=1}^n \sum_{j=1}^n A_{i,j} e^{\nu_i^{(1)}} e^{\nu_j^{(2)}} (w_i^{(1)} + w_j^{(2)})^2 \ge 0,$$
(71)

For each vector w in the null space of $\nabla^2 \mathbf{f}$, from (71), w satisfies $\langle w, \nabla^2 \mathbf{f}(\nu) w \rangle = 0$, which implies

$$w_i^{(1)} + w_j^{(2)} = 0$$
 for all $A_{i,j} > 0.$ (72)

Since A has support, then $\{A_{i,\sigma(i)} : i = 1, 2, ..., n\}$ are all positive for some permutation σ . Since $A_{i,\sigma(i)} > 0$, then any vector w in the null space of $\nabla^2 \mathbf{f}$ satisfies $w_i^{(1)} = -w_{\sigma(i)}^{(2)}$ and has the form

$$w := [w^{(1)}; w^{(2)}] = [w_1, w_2, \dots, w_n, -w_{\sigma^{-1}(1)}, \dots, -w_{\sigma^{-1}(n)}]^\top.$$
(73)

Clearly, $\langle w^{(1)}, \mathbb{1}_n \rangle + \langle w^{(2)}, \mathbb{1}_n \rangle = 0$ holds. Thus, we have the orthogonality between $-\nabla \mathbf{f}(\nu_k)$ and the null space of $\nabla^2 \mathbf{f}(\nu)$. Indeed,

$$\langle w, \nabla \mathbf{f}(\nu_k) \rangle = w^{(1)^{\top}} (\mathbf{B}(\nu_k) - I_n) \mathbb{1}_n + w^{(2)^{\top}} (\mathbf{B}(\nu_k) - I_n)^{\top} \mathbb{1}_n$$
(74)

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} (w_i^{(1)} + w_j^{(2)}) A_{i,j} e^{\nu_i^{(1)}} e^{\nu_j^{(2)}} = 0,$$
(75)

where we used (72). Hence, $-\nabla \mathbf{f}(\nu_k)$ lies in the range of $\nabla^2 \mathbf{f}$, which verifies that the system in (69) is consistent. Finally, we obtain (70) according to the positive semi-definite property in (71) and the following observation. Since $\nabla \mathbf{f}(\nu_k)$ is orthogonal to the null space of $\nabla^2 \mathbf{f}(\nu_k)$, then $\nabla \mathbf{f}(\nu_k)$ is orthogonal to the null space of $(\nabla^2 \mathbf{f}(\nu_k))^{\dagger}$.

Since $\nabla^2 \mathbf{f}(\nu_k)(\nu_{k+1}-\nu_k) = -\nabla \mathbf{f}(\nu_k)$ is consistent, the Newton iterations in (68) can be employed to find ν with $\nabla \mathbf{f}(\nu) = 0$, e.g., the conjugate gradient method [BCLW17]. Note that the squared Newton decrement in (70) can be interpreted as the directional derivative of \mathbf{f} in the direction of u,

$$-\langle u, \nabla^2 \mathbf{f}(\nu_k) u \rangle = \langle \nabla \mathbf{f}(\nu_k), u \rangle = \frac{d}{d\alpha} \mathbf{f}(\nu_k + \alpha u)|_{\alpha = 0}.$$
 (76)

Thanks to (70), when $\nabla \mathbf{f}(\nu_k) \neq 0$, the step size $\alpha > 0$ can be chosen properly to decrease the objective **f**.

Remark 3.3. Consider the application in SNNE, i.e., the balancing in (27). Note that the Hessian $\nabla^2 \mathbf{f}$ is exactly the Schur complement matrix \widetilde{M} described in (21). When $\mathbb{T}(x) = \exp(-t\mathbb{T}(c - M^{\top}\nu))$ heads to an optimal permutation with $t \to \infty$, the Hessian matrix will easily undergo a rank-reduction process. Hence, using a regularized Hessian matrix as in (23) is suggested in empirical algorithms for (68).

3.3 Logarithmic barrier functions(LB) based matrix balancing

We provide another Newton method, called Logarithmic barrier (LB) based matrix balancing, to compute scaling vectors of matrix balancing. The LB iterations will be stated in (84). The introduction can shed light on convergence of Knight-Ruiz algorithm. Consider a nonnegative matrix A. Define \tilde{A} as in (54). Consider the minimization of \mathbf{g} ,

$$\min_{\zeta>0} \{ \mathbf{g}(\zeta) = \frac{1}{2} \zeta^{\top} \widetilde{A} \zeta - \mathbb{1}_{2n}^{\top} \log \zeta \}.$$
(77)

The objective function in (77) is identical to the function in (60), except for ν replaced with $\log \zeta$. In [MO68], the function **g** is employed to show the existence of matrix-scaling on a fully indecomposable matrix. In [KK92], authors proposed one path-following Newton algorithm, minimizing a sequence of sub-problems to scale a symmetric positive semi-definite matrix \widetilde{A} , so that convergence requirement of Newton iterates can be met in each sub-problem. Here, we propose a modified Newton method for the computation of matrix balancing for one positive matrix A.

Compute the gradient and the Hessian of \mathbf{g} ,

$$\nabla \mathbf{g} = \widetilde{A}\zeta - \zeta^{-1}, \ \nabla^2 \mathbf{g}(\zeta) = \widetilde{A} + \operatorname{diag}(\zeta^{-2}), \tag{78}$$

respectively. First, from (78), the Sinkhorn-Knopp balancing is the coordinate descent iteration of $\mathbf{g}(\zeta)$ with $\zeta = [\zeta^{(1)}; \zeta^{(2)}]$,

$$\zeta_{k+1}^{(1)} \leftarrow \arg\min_{\zeta^{(1)}} \mathbf{g}([\zeta^{(1)}; \zeta_k^{(2)}]), \ \zeta_{k+1}^{(2)} \leftarrow \arg\min_{\zeta^{(2)}} \mathbf{g}([\zeta_{k+1}^{(1)}; \zeta^{(2)}]).$$
(79)

Thus, SK balancing decreases the objective \mathbf{g} in (77). Second, suppose a minimizer ζ is an interior point in \mathbb{R}^{2n}_+ . Clearly, ζ is a root to (56), i.e., $\widetilde{A}\zeta = \zeta^{-1}$. Write $\zeta = \exp(\nu)$ component-wise with some vector ν . From (71), $\mathbf{g}(\exp(\nu)) = \mathbf{f}(\nu)$ is convex in ν and a local minimizer of \mathbf{g} is actually the global minimizer of \mathbf{g} . Let us employ one damped Newton iteration to reach the global minimizer, where step size α_k is selected to minimize $\mathbf{g}(\zeta_k - \alpha_k(\nabla^2 \mathbf{g}(\zeta_k))^{-1}\nabla \mathbf{g}(\zeta_k))$ in (77), for $k = 1, 2, 3, \ldots$,

$$\zeta_{k+1} = \zeta_k - \alpha_k (\nabla^2 \mathbf{g}(\zeta_k))^{-1} \nabla \mathbf{g}(\zeta_k) = \zeta_k - \alpha_k (\widetilde{A} + \operatorname{diag}(\zeta_k^{-2}))^{-1} (\widetilde{A}\zeta_k - \zeta_k^{-1})$$
(80)

$$= \zeta_k - \alpha_k \operatorname{diag}(\zeta_k) \left(\operatorname{diag}(\zeta_k)(\widetilde{A} + \operatorname{diag}(\zeta_k^{-2})) \operatorname{diag}(\zeta_k) \right)^{-1} \left(\zeta_k \odot (\widetilde{A}\zeta_k - \zeta_k^{-1}) \right)$$
(81)

$$= \zeta_k - \alpha_k \zeta_k \odot (I_{2n} + B_k)^{-1} (B_k \mathbb{1}_{2n} - \mathbb{1}_{2n}),$$
(82)

with

$$B_k := \operatorname{diag}(\zeta_k) \widetilde{A} \operatorname{diag}(\zeta_k). \tag{83}$$

Since the matrix $I_{2n} + B_k$ in (82) is not necessarily positive definite, the iteration in (82) is not globally convergent. Instead, consider a modified Newton iteration (called LB matrix balancing scheme)

$$\zeta_{k+1} = \zeta_k \odot \{ \mathbb{1}_{2n} - \alpha_k (C_k + B_k)^{\dagger} (B_k - I_{2n}) \mathbb{1}_{2n} \},$$
(84)

where I_{2n} in (82) is replaced with the positive diagonal matrix,

$$C_k = \operatorname{diag}(\mathbb{1}_{2n} \odot (B_k \mathbb{1}_{2n})). \tag{85}$$

Remark 3.4 (Safeguard parameter ϵ_+). We implement (84) as follows. For each k, compute B_k and C_k from (83, 85), and $u_k := -\zeta_k \odot (C_k + B_k)^{\dagger} (B_k - I_{2n}) \mathbb{1}_{2n}$. Use conjugate gradient to solve

$$y_k := (C_k + B_k)^{\dagger} (B_k - I_{2n}) \mathbb{1}_{2n}$$
(86)

from the consistent system,

$$(C_k + B_k)y_k = (B_k - I_{2n})\mathbb{1}_{2n}.$$
(87)

The step size α_k is chosen to ensure the decrease of \mathbf{g} and $\zeta_{k+1} = \zeta_k \odot (1 - \alpha_k y_k) > 0$. For $\zeta_{k+1} > 0$, we introduce a safeguard parameter $\epsilon_+ \in (0, 1)$ and α is chosen within $(0, y_{\max}^{-1}(1 - \epsilon_+)]$, where y_{\max} is

the largest positive entry of y_k . Indeed, $\zeta_{k+1} = \zeta_k + \alpha u_k = \zeta_k \odot (1 - \alpha y_k) \ge \zeta_k \epsilon_+ > 0$.

In the following, we shall discuss the wellposeness of LB and show the step size of LB tending to 1 near an optimal solution.

3.3.1 Well-definedness of LB in (84)

The following proposition shows the well-definedness of $(C_k + B_k)^{\dagger}(B_k - I_{2n})\mathbb{1}_{2n}$ in (84). Also, we calculate the directional derivative of **g** in the direction of

$$u_k := -\zeta_k \odot (C_k + B_k)^{\dagger} (B_k - I_{2n}) \mathbb{1}_{2n}$$
(88)

with $||u_k|| > 0$, which sheds some light on the convergence of this Newton method,

$$\frac{d}{d\alpha}\mathbf{g}(\zeta_k + \alpha u_k)|_{\alpha=0} = \langle \nabla \mathbf{g}(\zeta_k), u_k \rangle \tag{89}$$

$$= -\langle (B_k - I_{2n})\mathbb{1}_{2n}, (C_k + B_k)^{\dagger} (B_k - I_{2n})\mathbb{1}_{2n} \rangle < 0.$$
(90)

In the following, we shall verify the calculation in (90). We introduce $H(\zeta)$ in (94) to investigate the null space of $B_k + C_k$. Note that $H(\zeta_k) = C_k + B_k$.

Proposition 3.5. Consider one matrix $A \in \mathbb{R}^{n \times n}$, which is nonnegative and has support. Let \widetilde{A} be given in (54) and $B = \operatorname{diag}(\zeta)\widetilde{A}\operatorname{diag}(\zeta)$. Let $C = \operatorname{diag}(B\mathbb{1}_{2n})$. Then C + B is symmetric and positive semi-definite and the system

$$(C+B)y = (B-I_{2n})\mathbb{1}_{2n} \tag{91}$$

is consistent. In addition, introduce the null space of C + B,

$$\mathcal{N} := \{ w = [w^{(1)}; w^{(2)}] : w_i^{(1)} + w_j^{(2)} = 0, \ \forall (i,j) \ with \ A_{i,j} > 0 \}.$$
(92)

For any positive vector $\zeta \in \mathbb{R}^{2n}$ and for any null vector $w \in \mathcal{N}$, the function **g** takes a constant value, as $\zeta \to \zeta \odot \exp(w)$, *i.e.*,

$$\mathbf{g}(\zeta \odot \exp(w)) = \mathbf{g}(\zeta). \tag{93}$$

Introduce

$$H(\zeta) = \operatorname{diag}(\zeta \odot (\widetilde{A}\zeta)) + \operatorname{diag}(\zeta)\widetilde{A}\operatorname{diag}(\zeta).$$
(94)

Then \mathcal{N} is the null space of $H(\zeta)$ for any positive vector ζ .

Proof. By Gershgorin circle theorem, the symmetric matrix $C + B \succeq 0$ is diagonally dominant and thus is a positive semi-definite matrix. Actually, for each vector $w = [w^{(1)}; w^{(2)}] \in \mathbb{R}^{2n}$,

$$\langle w, (C+B)w \rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{i,j} \zeta_i^{(1)} \zeta_j^{(2)} (w_i^{(1)} + w_j^{(2)})^2 \ge 0.$$
(95)

Hence, each null vector w of C + B satisfies

$$w_i^{(1)} + w_j^{(2)} = 0 \text{ for all } A_{i,j} > 0,$$
 (96)

which justifies (92). Since A has support, then for some permutation σ , we have $A_{i,\sigma(i)} > 0$ for $i = 1, \ldots, n$. Hence, (96) implies

$$\sum_{i=1}^{n} w_i^{(1)} + \sum_{j=1}^{n} w_j^{(2)} = \sum_{i=1}^{n} w_i^{(1)} + \sum_{i=1}^{n} w_{\sigma(i)}^{(2)} = 0.$$
(97)

Next, we show that $(B - I_{2n})\mathbb{1}_{2n}$ lies in the range of (C + B). Indeed, for each null vector w, using (96) and (97), we have $(B - I_{2n})\mathbb{1}_{2n}$ is orthogonal to the null space of $(C + B)^{\top} = C + B$, i.e.,

$$\langle w, (B - I_{2n}) \mathbb{1}_{2n} \rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} (w_i^{(1)} + w_j^{(2)}) \zeta_i^{(1)} \zeta_j^{(2)} A_{i,j} - (\sum_{i=1}^{n} w_i^{(1)} + \sum_{j=1}^{n} w_j^{(2)}) = 0.$$
(98)

The above orthogonality arguments also implies (90). Again from (96) and (97), we have

$$\mathbf{g}(\zeta \odot \exp(w)) = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{i,j} \zeta_i^{(1)} \zeta_j^{(2)} \exp(w_i^{(1)} + w_j^{(2)}) - \langle \mathbb{1}_{2n}, \log \zeta \rangle - \langle \mathbb{1}_{2n}, w \rangle$$
(99)

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} A_{i,j} \zeta_{i}^{(1)} \zeta_{j}^{(2)} - \langle \mathbb{1}_{2n}, \log \zeta \rangle = \mathbf{g}(\zeta).$$
(100)

Finally, observe that $\langle w, H(\zeta)w \rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{i,j} \zeta_{i}^{(1)} \zeta_{j}^{(2)} (w_{i}^{(1)} + w_{j}^{(2)})^{2} = 0$ if and only if $w \in \mathcal{N}$. Thus, \mathcal{N} is the null space of $H(\zeta)$ for any positive vector ζ .

3.3.2 Relation between KR and LB

First, we make one observation.

Remark 3.6 (KR method is a special case with $\alpha_k = 1$). Note that the LB method in (84) with $\alpha_k = 1$ coincides with the algorithm proposed by Knight and Ruiz in (57). As k increases, the objective values $\mathbf{g}(\zeta_k)$ decrease monotonically. As B_k tends to be a doubly stochastic matrix, we have $C_k = \text{diag}(B_k \mathbb{1}_{2n}) \rightarrow I_{2n}$ and the LB method reduces to Newton's method in (57), i.e., KR method.

In the following, we demonstrate that the step size α_k of LB is 1 for sufficiently large k. To proceed, we start with some boundedness related to the sequence $\{\zeta_k : \mathbf{g}(\zeta_k) \leq c_0, k = 1, 2, ...\}$ under total support assumption on A. For notation simplicity, we drop the subscript k.

Proposition 3.7. Suppose $A \in \mathbb{R}^{n \times n}$ has total support. Let $\Sigma := \{(i, j) : A_{i,j} > 0\}$. Let δ be a positive lower bound for $\{A_{i,j} : (i, j) \in \Sigma\}$. Fix some $c_0 \in \mathbb{R}$. Let $\zeta = [\zeta^{(1)}, \zeta^{(2)}]$ be a positive vector in the c_0 -sublevel set of \mathbf{g} , i.e., $\mathbf{g}(\zeta) \leq c_0$. Then $\{\zeta_i^{(1)}\zeta_j^{(2)} : (i, j) \in \Sigma\}$ are bounded below by

$$\exp(-c_0 + (n-1)(1+\log\delta)) \tag{101}$$

and bounded above by $% \label{eq:logistical} \left(\begin{array}{c} \mbox{and} \ \mbox{bounded} \ \mbox{boundedddddd\ \mbox{bounded} \ \mbox{bounded} \ \mb$

$$\max(\delta^{-1}(c_0 - (n-1)(1 + \log \delta)), 1).$$
(102)

In particular, for any ζ with $\mathbf{g}(\zeta) \leq c_0$, $\|(\zeta^{(1)}\zeta^{(2)^{\top}}) \odot \mathbb{1}_{\Sigma}\|$ is bounded above by some constant only depending on c_0 and δ .

Proof. Fix one entry $A_{i_1,j_1} > 0$. By assumption, A has total support, and thus (i_1, j_1) lies on some diagonal $\{(i, \sigma(i)) : i \in \{1, 2, ..., n\}\}$. Then

$$\sum_{i=1}^{n} \{A_{i,\sigma(i)}\zeta_{i}^{(1)}\zeta_{\sigma(i)}^{(2)} - \log(\zeta_{i}^{(1)}\zeta_{\sigma(i)}^{(2)})\} \le \mathbf{g}(\zeta) = \langle \zeta^{(1)}, A\zeta^{(2)} \rangle - \langle \mathbb{1}_{2n}, \log \zeta \rangle \le c_{0}.$$
(103)

By convexity, the following inequality holds for each a > 0,

$$\min_{x \ge 0} (ax - \log x) \ge 1 + \log a.$$
(104)

Applying (104) to the right hand side of (103) for those $i \neq i_1$, we have

$$\sum_{i=1}^{n} \{A_{i,\sigma(i)}\zeta_{i}^{(1)}\zeta_{\sigma(i)}^{(2)} - \log(\zeta_{i}^{(1)}\zeta_{\sigma(i)}^{(2)})\} \ge A_{i_{1},j_{1}}\zeta_{i_{1}}^{(1)}\zeta_{j_{1}}^{(2)} - \log(\zeta_{i_{1}}^{(1)}\zeta_{j_{1}}^{(2)}) + (n-1)(1+\log\delta).$$
(105)

Together with (103), dropping the positive term $A_{i_1,j_1}\zeta_{i_1}^{(1)}\zeta_{j_1}^{(2)}$ in (105), we have (101). Likewise, for an upper bound, when $\zeta_{i_1}^{(1)}\zeta_{j_1}^{(2)} \ge 1$, we can drop $-\log(\zeta_{i_1}^{(1)}\zeta_{j_1}^{(2)})$ in (105), which yields the upper bound in (102).

Remark 3.8. Let $\mathbb{1}_{\Sigma} := (A > 0)$. When A has total support, then $\zeta_k = [\zeta_k^{(1)}, \zeta_k^{(2)}]$ from (84) generates a bounded matrix $(\zeta_k^{(1)} \zeta_k^{(2)^{\top}}) \odot \mathbb{1}_{\Sigma} \in \mathbb{R}^{n \times n}$. Express the k-th iterate ζ_k as $\zeta_k = \exp(\nu_k)$ with $\nu_k = [\nu_k^{(1)}; \nu_k^{(2)}]$. Introduce a linear transform \mathbb{B} ,

$$\mathbb{B}(\nu_k) := \mathbb{1}_{\Sigma} \odot \log(\zeta_k^{(1)} \zeta_k^{(2)^{\top}}) = \mathbb{1}_{\Sigma} \odot \mathbb{T}(M^{\top} \nu_k) = \mathbb{T}((\mathbb{T}^{-1}(\mathbb{1}_{\Sigma})) \odot M^{\top} \nu_k).$$
(106)

From Prop. 3.7, the null space of \mathbb{B} is the null space \mathcal{N} in (92), i.e.,

$$\mathcal{N} = \{ w : \mathbb{T}^{-1}(\mathbb{1}_{\Sigma}) \odot M^{\top} w = 0 \} = \{ w : A \odot (w^{(1)} \mathbb{1}_{n}^{\top} + \mathbb{1}_{n} w^{(2)^{\top}}) = 0 \}.$$
(107)

Let $P : \mathbb{R}^{2n} \to \mathbb{R}^{2n}$ be the orthogonal projection with kernel \mathcal{N} . Let m be the smallest singular value of \mathbb{B} . Then $\|\mathbb{B}\nu_k\| \ge m\|P\nu_k\|$. Hence, the boundedness $(\zeta_k^{(1)}\zeta_k^{(2)^\top}) \odot \mathbb{1}_{\Sigma}$ actually indicates the boundedness of $\{\|P\nu_k\| : k = 1, 2, 3, ...\}$, when $\{\nu_k\}$ and $\{\zeta_k\}$ are chosen to minimize $\mathbf{f}(\nu)$ or $\mathbf{g}(\zeta)$, respective. This justifies the norm assumption required in Theorem 2.

The following theorem states that LB iterates are exactly KR iterations, when k is sufficiently large. Since the proof is lengthy, we place it in the appendix.

Theorem 3. Suppose that $A \in \mathbb{R}^{n \times n}$ has total support. For k sufficiently large, the step size α_k in the LB iteration is 1.

3.4 Stability issues in practical algorithms

When we balance a sequence of matrices with t increasing, the norm of these scaling vectors will increase synchronously. Without careful numerical treatment, large numerical errors can easily occur in KR, NE and LB algorithms. Two techniques proposed in the Stabilized Scaling algorithms [Sch19] will be employed in our simulation studies of KR, NE and LB algorithms.

In the application of optimal transport, we are interested in balancing a sequence of matrices

$$A = \exp(-t\mathbb{T}(c)) \tag{108}$$

for a sequence of t-sequence, i.e., $\operatorname{diag}(\zeta^{(1)})A\operatorname{diag}(\zeta^{(2)})$ is doubly stochastic under some scaling vectors $\zeta^{(1)}, \zeta^{(2)}$. The first technique is that to avoid the numerical inaccuracy caused by the large entries in scaling vectors, we should execute matrix balancing algorithms in the Log-Domain. For instance, in the LB method, we shall avoid computing/storing $\zeta^{(1)}, \zeta^{(2)}$ in matrix balancing algorithms. Instead, by expressing $\zeta^{(1)}, \zeta^{(2)}$ as $\zeta^{(1)} = \exp(t\nu^{(1)})$ and $\zeta^{(2)} = \exp(t\nu^{(2)})$ for some $\nu = [\nu^{(1)}, \nu^{(2)}]$, we should conduct matrix balancing in terms of $\nu^{(1)}$ and $\nu^{(2)}$. Hence, the LB iteration in (84) should be rewritten as

$$\nu_{k+1} = \nu_k + t^{-1} \log(\mathbb{1}_{2n} - \alpha_k (C_k + B_k)^{\dagger} (B_k - I_{2n}) \mathbb{1}_{2n}), \tag{109}$$

and $C_k + B_k$ can be expressed as

$$C_k + B_k = \begin{pmatrix} \operatorname{diag}(\mathbb{1}_{\Sigma} \odot \exp(-t\mathbb{T}(c - M^{\top}\nu_k)))\mathbb{1}_n & \mathbb{1}_{\Sigma} \odot \exp(-t\mathbb{T}(c - M^{\top}\nu_k)) \\ (\mathbb{1}_{\Sigma} \odot \exp(-t\mathbb{T}(c - M^{\top}\nu_k)))^{\top} & \operatorname{diag}((\mathbb{1}_{\Sigma} \odot \exp(-t\mathbb{T}(c - M^{\top}\nu_k)))^{\top}\mathbb{1}_n) \end{pmatrix}.$$
(110)

The second technique is to use μ -translation to reduce numerical errors in matrix balancing computation. Suppose the scaling vectors $\{\exp(t\nu^{(1)}), \exp(t\nu^{(2)})\}$ for the squared matrix $A = \exp(-t\mathbb{T}(c)) \in \mathbb{R}^{n,n}$ is available. Then the squared (shifted) matrix

$$\exp(-t(\mathbb{T}(c) - M^{\top}\mu)) = \operatorname{diag}(t\mu^{(1)})A\operatorname{diag}(t\mu^{(2)})$$
(111)

can be balanced by translated scaling vectors $\{\exp(t(\nu^{(1)} - \mu^{(1)})), \exp(t(\nu^{(2)} - \mu^{(2)}))\}$. How should we choose $\{\mu^{(1)}, \mu^{(2)}\}$? Suppose $\exp(-t_{k-1}\mathbb{T}(c))$ can be balanced by scaling vectors $\{\exp(t_{k-1}\nu^{(1)}), \exp(t_{k-1}\nu^{(2)})\}$. When t_{k-1} is sufficiently large, $\{\exp(t_k\nu^{(1)}), \exp(t_k\nu^{(2)})\}$ provides a good approximation for scaling vectors of $\exp(-t_k\mathbb{T}(c))$. Thus, one good empirical choice is $\mu^{(1)} = \nu^{(1)}$ and $\mu^{(2)} = \nu^{(2)}$. Once the scaling vectors $\{\exp(t_k\xi^{(1)}), \exp(t_k\xi^{(2)})\}$ of the shifted matrix

$$\exp(-t_k \mathbb{T}(c - M^{\top} \nu)) \tag{112}$$

are computed, we know that the original matrix $\exp(-t_k \mathbb{T}(c))$ in (108) can be balanced by scaling vectors $\{\exp(t_k(\nu^{(1)} + \xi^{(1)})), \exp(t_k(\nu^{(2)} + \xi^{(2)}))\}$. In summary, we have the following algorithm for the problem in (108).

Algorithm 3.9. Input: a matrix $\mathbb{T}(c) \in \mathbb{R}^{n \times n}$ and a sequence $t_1, t_2, \ldots, t_{\max}$ in \mathbb{R} .

- Initialize $\nu_0 = 0_{2n}$. For $k = 1, 2, \ldots, k_{\text{max}}$, repeat the following two steps:
- Compute a scaling vector $\exp(t\mu) \in \mathbb{R}^{2n}$ which balances the matrix $\exp(-t_k(\mathbb{T}(c) M^{\top}\nu_{k-1}))$.
- Update $\nu_k = \nu_{k-1} + \mu \in \mathbb{R}^{2n}$.

Output: $\nu_{k_{\max}}$. Here, the vector $\exp(t_{\max}\nu_{k_{\max}})$ balances the matrix $\exp(-t_{\max}\mathbb{T}(c))$.

4 Numerical simulations

We provide three experiments in the section: (i) Comparison of matrix balancing schemes; (ii)Comparison experiments of matrix balancing in solving discrete optimal transport; (iii) Application of sparse support algorithms on large data-sets.

4.1 Matrix balancing

4.1.1 Comparison in matrix balancing

We compare four matrix balancing methods, including

- Sinkhorn-Knopp algorithm(SK) in (53);
- three Newton method based algorithms:
 - Knight-Ruiz method(KR) in (57);
 - Negative entropy method(NE) in (68);
 - Logarithmic barrier method(LB) in (84).

We select three matrices, $A = \exp(-magic(20)/20)$ of size 20×20 , $A = \exp(-magic(50)/20)$ of size 50×50 , and $A = \exp(-magic(200)/50)$ of size 200×200 . Here magic(n) produces an $n \times n$ matrix from the integers $1, 2, \ldots, n^2$ with with equal row/column/diagonal sums. See the top row of Fig. 1 for the pattern visualization of matrices magic(20), magic(50) and magic(200). At the k-th iteration, let $B_k := \operatorname{diag}(\zeta_k^{(1)})A\operatorname{diag}(\zeta_k^{(2)})$ be the matrix corresponding to scaling vectors $\{\zeta_k^{(1)}, \zeta_k^{(2)}\}$ produced from matrix balancing algorithms. Consider the performance metric to evaluate the matrix balancing error:

$$Error := \|B_k \mathbb{1}_n - \mathbb{1}_n\|_1 + \|B_k^{\dagger} \mathbb{1}_n - \mathbb{1}_n\|_1.$$
(113)

- First, we start with the same initial vector $\mathbb{1}_n$ in the four methods. Results are reported in Figure 2, where KR empirically gives very fast convergence in the perspective of CPU time. Sinkhorn-Knopp algorithm, one popular algorithm, typically requires more iterations than Newton methods. However, thanks to its low complexity in each iteration, SK can produce acceptable results economically. For instance, as shown in $A = \exp(-magic(20)/20)$ and $A = \exp(-magic(200)/50)$, SK reaches a solution with error less than 10^{-2} , much faster than NE and LB. On the other hand, SK has very poor convergence in handling $A = \exp(-magic(50)/20)$. This case with n = 50 is actually a challenging problem. Optimal scaling vectors $\zeta^{(1)}, \zeta^{(2)}$ have norm both greater than 10^{12} , which suggest that $\exp(-magic(50)/20)$ nearly does not have total support. Under the circumstance, all Newton methods give relatively slow convergence.
- Second, we further examine the case $A = \exp(-magic(50)/20)$ from the framework of negative entropic barrier functions. Consider a sequence of matrices $\exp(-t \cdot magic(50))$ with t = 1/160, 1/80, 1/40 and 1/20, respectively. The CPU time of these balancing tasks is reported



Figure 1: (Left to right subfigures in the top row show matrices magic(20), magic(50) and magic(200), respectively. Left and right subfigures in the bottom row show magic(1950) and magic(2500), respectively.

	$\exp(-$	magic(50)	$() \cdot t)$	
t relue	NE	LB	KR	SK
<i>t</i> value	(s)	(s)	(s)	(s)
1/160	0.0032	0.0018	0.0007	0.025
1/80	0.0047	0.0021	0.0008	0.075
1/40	0.0071	0.0039	0.0013	0.140
1/20	0.0074	0.0042	0.0025	0.939

Table 1: Computational time (sec) in balancing $\exp(-magic(50) \cdot t)$ under tolerance 10^{-5} .

in Table 1. Matrix balancing task with small t is easier than those tasks with large t. For t = 1/160, 1/80, 1/40, 1/20, the geometric mean of the norm of the scaling vectors is

$$\|\zeta^{(1)}\|^{1/2} \|\zeta^{(2)}\|^{1/2} = 2.31 \times 10^1, \ 8.05 \times 10^2, \ 1.61 \times 10^6, \ 1.08 \times 10^{13}, \tag{114}$$

respectively.¹ From Remark 3.8, the norm growth of scaling vectors reflects that the matrices to be balanced nearly do not have total support. In addition, we examine the scaling vectors

$$\zeta^{(1)} = \exp(t\nu^{(1)}), \zeta^{(2)} = \exp(t\nu^{(2)}), \tag{115}$$

by plotting those entries of dual vectors $\nu^{(1)}$ and $\nu^{(2)}$ in Fig. 3. Observe the similarity among these vectors $\nu^{(1)}$ and vectors $\nu^{(2)}$. Fast convergence of Newton methods relies on the proximity of the initialization to the attractive basin. Thanks to the similarity, we can speed up these Newton methods, when the optimal scaling vectors of matrices with previous t are employed as warm starts. Notice that the CPU time with t = 1/20 is improved significantly, compared with CPU time reported in Fig. 2.

¹As one reference, $\|\zeta^{(1)}\|^{1/2} \|\zeta^{(2)}\|^{1/2}$ is 1.663 and 137.8 for the problems $\exp(-magic(20)/20)$ and $\exp(-magic(200)/20)$, respectively.



Figure 2: Comparison of SK with other Newton based matrix balancings, n = 20(left), n = 50(middle), and n = 200 (right). The performance metric is (113).



Figure 3: Top: Vectors $\nu^{(1)}$ and $\nu^{(2)}$ of (115) in balancing $\exp(-magic(50) \cdot t)$. Bottom: Vectors $\nu^{(1)}$ and $\nu^{(2)}$ in optimal transport with $\mathbb{T}(c) = magic(1950)$.

4.1.2 Comparison in solving optimal transport

We demonstrate the application of matrix balancing algorithms in solving optimal transport along a central path for $t = t_k = t_0 \eta^k$, $k = 0, 1, 2, ..., t_{\text{max}}$. We evaluate matrix balancing algorithms in handling the cases with assignment matrix

- $\mathbb{T}(c) = magic(2500);$
- $\mathbb{T}(c) = magic(1950);$
- $\mathbb{T}(c)$ has n^2 entries $\{c_{j,k} = \|y_j z_k\|^2 : i, j = 1, ..., n\}$, where $\{y_k\}_{k=1}^n$ is one TLC point-set and and $\{z_k\}_{k=1}^n$ is one translated FRC point-set, n = 254.

Perform matrix balancing of a sequence of matrices for a few positive values $t = t_k$ until the relative duality gap ϵ is met, i.e,

$$\{\exp(-t_i \mathbb{T}(c)) : t_1 < t_2 < \dots < t_{\max}\}.$$
(116)

Let x_{opt} be an optimal primal vector. For each algorithm, we report the computation time, when relative duality gap falls within a given tolerance level ϵ ,

$$\langle c, x_{opt} \rangle^{-1} (\langle c, x \rangle - \langle \nu, \mathbb{1}_{2n} \rangle) \le \epsilon.$$
 (117)

Table 2-4 report the CPU time and the corresponding t_k value for various tolerance level ϵ . Notice that when identical sequences of t_k are reported, identical sequence of matrices are balanced in these methods. Consider a fixed matrix balancing tolerance $\epsilon_{MB} = 10^{-5}\sqrt{n}$ as the stopping criterion. This criterion ensures that the gradient has small norm, $\|\nabla \mathbf{f}(\nu)\| \leq \epsilon_{MB}$, see (65). Experiment results show that all Newton methods work quite well in the three problems. In particular, KR consistently gives the fastest convergence among these Newton methods. However, a winner between Newton methods and SK usually depends on the difficulty of the problem itself. Observe the pattern similarity between magic(200) and magic(2500) and observe the pattern similarity between magic(50) and magic(1950)from Fig. 1. For the problem magic(2500), which is relatively easy (compared with magic(1950)), SK is a fast algorithm, which produces acceptable results, much faster than NE and LB as shown in magic(200). On the other hand, facing the challenging problem magic(1950), SK fails to produce acceptable results within 5000 seconds. As a result, we can see the similarity of the dual vectors $\nu^{(1)}, \nu^{(2)}$ in Fig. 3. As in magic(50) and magic(1950), entries of dual vectors in a point-set matching problem actually vary a lot. From this viewpoint, it is not so surprising that SK has the worst convergence in solving the point set matching problem, shown in Table 4.

			n	nagic(2500))			
6	NE		LB		KR		SK	
L	time	t_k	time	t_k	time	t_k	time	t_k
1e - 1	22.35	291.9	43.73	291.9	13.77	291.9	2.99	291.9
1e - 2	32.10	3325	62.21	3325	19.15	3325	4.42	3325
1e - 3	45.96	37876	80.46	37876	24.56	37876	6.15	37876
1e - 4	51.55	287627	94.45	287627	28.73	287627	7.00	287627
1e - 5	116.87	2675044	108.32	4914369	34.00	3276247	8.17	3276247

Table 2: Computational time(sec) in solving optimal transport with $\mathbb{T}(c) = magic(2500)$.

4.2 Rigid-motion estimation

One big advantage of SNNE over primal-dual methods is that SNNE updates multiplier vectors solely along the increase of t, i.e., no need to store/pass x between sub-problems. The memory requirement in SNNE can be much less than that in primal-dual methods, if the active support set is properly handled in large-scale problems. The following two experiments demonstrate the effectiveness of SNNE in handling large-scale problems. In the first study, we provide one comparison between SNNE, SNNE-sparse

				magic(198)	50)			
c	NE]	LB		KR	SK	
C	time	t_k	time	t_k	time	t_k	time	t_k
1e - 1	25.29	437.9	66.42	437.9	15.30	437.9	11197	388.2
1e - 2	34.36	3325	96.84	3325	19.89	3325	11626	4.42×10^3
1e - 3	46.63	37877	133.38	37877	25.71	37877	-	-
1e - 4	65.09	287627	190.35	287627	41.22	287627	-	-
1e - 5	129.21	2184164	2796	1531812	389.12	2184164	-	-

Table 3: Computational time(sec) in solving optimal transport with $\mathbb{T}(c) = magic(1950)$.

	Lung branch points $(n = 254)$							
C	NE		LB		KR		SK	
C	time	t_k	time	t_k	time	t_k	time	t_k
1e - 1	0.14	86.5	0.32	86.5	0.11	86.5	4.94	70.6
1e-2	0.22	4.37×10^2	0.40	4.37×10^2	0.13	4.37×10^2	25.26	509.8
1e-3	0.39	1.478×10^{3}	0.64	1.478×10^{3}	0.28	1.478×10^{3}	200.46	1.348×10^{3}
1e-4	1.10	4.988×10^{3}	5.19	4.988×10^{3}	0.66	4.988×10^{3}	851.94	5.247×10^3
1e-5	1.25	2.5251×10^4	5.89	2.5251×10^4	0.80	2.5251×10^4	859.51	2.1621×10^4

Table 4: Computational time(sec) in solving optimal transport with $c = L^2$ -distance assignment.

with primal-dual methods, which are popularly used in solving linear programming. Here, we consider two primal-dual methods: Mehrotra predictor-corrector method, which is one widely-used primal-dual interior point method [Meh92], and one commercial software solver, Gurobi, where the algorithm method is chosen to be the barrier method. In the first study, we actually solve a number of optimal transport problems. For the second study, we demonstrate the flexibility of the entropic regularization. We apply entropic regularization, but take t as the outer loop variable to bypass the multiple optimal transport problems. The algorithms SNNE-t and SNNE-sparse are developed in this framework to optimize the computational time.

We present a rigid motion experiment on a three-dimensional teapot point cloud consisting of 41472 points. We subsample 1000/2500/5000 point-sets $\{y_1, \ldots, y_n\}$ from the teapot point cloud. Select one orthogonal matrix $Q \in \mathbb{R}^{3\times 3}$, and generate another set of point-sets, $\{z_i = Qy_i : i = 1, \ldots, n\}$, as shown in Figure 4. For simplicity, $\{y_i\}$ is shifted so that $\sum_{i=1}^n y_i = 0$. Introducing a user-defined parameter $\eta > 0$, we estimate $Q \in \mathbb{R}^{3\times 3}$ and $\mathbb{T}(x) \in \Pi_n$ from the minimization,

$$\min_{Q} \min_{x} \left\{ \mathbb{F}(Q, x) := \langle \mathbf{c}(Q), x \rangle + \eta \| Q - I_3 \|_F^2 \right\},\tag{118}$$

where the assignment $\mathbf{c}(Q)$ is a function of Q with $\mathbb{T}(\mathbf{c}(Q))_{i,j} = ||y_i - Qz_j||^2$. We can apply optimal transport for general non-rigid motion problems via introducing regulation terms for splines. For instance, see [CR00, GTY04, Che11a].

Here is one *naive* algorithm, consisting of repeating the estimations of Q and x:

- Fix Q. Estimate x with $\mathbb{T}(x) \in \Pi_n$, which is one optimal transport.
- Fix x. Solve Q from the least squares problem,

$$\min_{Q} \{ \mathbb{F}(Q, x) = \sum_{i,j=1}^{n} \langle (y_i - Qz_j), x_{i,j}(y_i - Qz_j) \rangle + \eta \langle Q - I, Q - I \rangle \}.$$
(119)

From the SVD property, an optimal matrix is $Q = UV^{\top}$, where U, V are unitary matrices in the SVD,

$$UDV^{\top} = \sum_{i,j=1}^{n} \{ x_{i,j} y_i z_j^{\top} + \eta I \}.$$
 (120)

The performance metric is given by

$$error := \langle \mathbf{c}(Q), x \rangle \ge 0. \tag{121}$$

Note that error = 0 if and only if $y_i = Qz_j$ holds for all $x_{i,j} > 0$.

For a fair comparison, we use SNNE, Mehrotra primal-dual method(PD) and Gurobi solver to solve optimal transport minimizer $\mathbb{T}(x)$ after each Q-update. Table 5 reports the computational time of SNNE, Mehrotra primal-dual method(PD) and Gurobi optimization software. We stop algorithms when *error* reaches 10^{-4} . Figure 5 shows the desired small error under PD, Gurobi and SNNE, which indicates the successful reconstruction of x and Q in the cases n = 2500 and n = 5000. As expected, when nincreases, the computational time increases accordingly. The computational time of PD is approximately proportional to n^3 , while the computational time of SNNE or Gurobi is approximately proportional to n^2 . Clearly, either Gurobi optimization software or Mehrotra predictor-corrector method can deliver an optimal solution of optimal transport in (3) very fast, when the cardinality n does not exceed 1000. However, due to its advantage in low memory requirement, the inferior performance of SNNE becomes less apparent in the case n = 2500 and n = 5000. See Figure 5.

4.2.1 SNNE-t and SNNE-sparse

In SNNE, after each Q-update, a sequence of matrices are balanced to generate one approximate optimal transport minimizer for each assignment matrix $\mathbf{c}(Q)$. Balancing these matrices along multiple paths actually makes SNNE very inefficient. To alleviate the difficulty, we introduce the entropic regularization to (118) to estimate (Q, x) along "one" inexact minimizer path associated with a sequence $\{t = t_1, \ldots, t = t_{\max}\}$,

$$\min_{Q} \min_{\Sigma} \min_{x} \left\{ \mathbb{F}_{t}(Q, x, \Sigma) := \langle \mathbf{c}(Q), x \rangle - t^{-1} \langle \mathbb{T}^{-1}(\mathbb{1}_{\Sigma}), \log x \rangle + \eta \|Q - I_{3}\|_{F}^{2} \right\}.$$
(122)

At each t, we execute the following block coordinate steps to approximate the minimizer (Q, x).

- Fixing Q, use matrix balancing to compute an optimal $\mathbb{T}(x)$, i.e., find ν to balance the matrix $\exp(-t(c(Q) M^{\top}\nu))$. Use ν to update Σ .
- Fixing x, we update Q by SVD computation in (120).

The convergence to the exact minimizer (Q, x) requires a sufficient number of these block coordinate descent steps. (See Prop. 2.7.1 [Ber03].) As t gets sufficiently large, (Q, x) in (122) is expected to approach one minimizer in (118). We call the new algorithm solving (122) along one t-path as **SNNE-t**. Note that the major difference from (118) is that the parameter t in (122) is an outer loop variable. Results are reported in Table 5. Thanks to bypassing multiple optimal transport problems, SNNE-t actually consumes much less computation time than previous algorithms.

Next, we implement SNNE-sparse to solve (122), where the support of x is dynamically updated reduce the memory load of SNNE-t. That is, x and Q are updated alternately with initialization Q = I. For each Q fixed, we compute x via one approximate multiplier vector ν_{ξ} subject to the approximate support set Σ_{ξ} for ξ in $\{1, 2, 3, \ldots, \xi_{\max}\}$, as in Alg. 2.5. To have a better control on sparsity of Σ_{ξ} in SNNE, we select a sparse parameter k = 20 to ensure an upper bound (2k+1)n for the cardinality of Σ_{ξ} . The result of SNNE-sparse is reported in Table 5 and Fig. 5. Clearly, the introduction of matrix sparsity together with the usage of one *t*-path greatly reduces the computational time of the implementation of SNNE-sparse. Here, $\xi_{\max} = 3$ is used. The heuristic choice of ξ_{\max} has a big influence on the whole computational time. When $\xi_{\max} = 2$ is used, the computation time can be further reduced. See the column of SNNE-sparse-2 in Table 5.

Remark 4.1 (Multi-scale similarity). Actually the multiplier vectors corresponding to different cardinality n resemble each other. The dual vector associated with coarser sampling can be used as one warm start to compute the dual vector associated with finer sampling. For instance, consider the application of SNNE on the problem with n = 2500 and n = 5000, respectively,

$$\min_{\mathbb{T}(x)\in\Pi_n} \langle c, x \rangle, \text{ with } \mathbb{T}(c)_{i,j} = \|y_i - z_j\|^2, i, j = 1, \dots, n.$$



Figure 4: Left: the teapot point set (41472 points). Middle: the point set $\{y_i : i = 1, ..., 5000\}$. Right: the point set $\{z_i : i = 1, ..., 5000\}$.



Figure 5: Computational time(sec) vs. error metric in (121) under SNNE methods and Primal-dual methods.

For the case n = 2500, let $Y' = \{y_1, \ldots, y_{2500}\}$ and $Z' = \{z_1, \ldots, z_{2500}\}$. Let $[\nu^{(1)'}, \nu^{(2)'}]$ be the multiplier vector in (25). For the case n = 5000, let $Y = \{y_1, \ldots, y_{5000}\}$ and $Z = \{z_1, \ldots, z_{5000}\}$. Let $[\nu^{(1)}, \nu^{(2)}]$ be the multiplier vector in (25). The color distribution in the top figures showing $(\nu^{(1)'}, \nu^{(2)'})$ resembles the color distribution in the bottom figures showing $(\nu^{(1)}, \nu^{(2)})$. Indeed, $\nu^{(1)} \approx \nu^{(1)'} + 160$ and $\nu^{(2)} \approx \nu^{(2)'} - 160$. (Here the shift is caused by the one-dimension null space of M.) Hence, we can employ $(\nu^{(1)'}, \nu^{(2)'})$ to produce a warm start $(\nu^{(1)}_{ini}, \nu^{(2)}_{ini})$ (satisfying KKT conditions in (7)) to initialize ν (which initializes Σ) in the problem with n = 5000. That is,

• let $\nu_{ini}^{(1)}$ be computed as follows: for $j = 1, \dots, 5000$

$$\nu_{ini}^{(1)}(j) = \max_{k} \{ \|y_j - z_k\|^2 - {\nu^{(2)}}'(k) : z_k \in Z' \}.$$
(123)

• Let $\nu_{ini}^{(2)}$ be computed as follows: for $k = 1, \dots, 5000$

$$\nu_{ini}^{(2)}(k) = \max_{j} \{ \|y_j - z_k\|^2 - \nu_{ini}^{(1)}(j) : y_j \in Y \}.$$
(124)

4.2.2 Support sets without total support

The following provides one comparison between the performance under sparse support sets given by $\Sigma = \Sigma' \cup \Sigma'' \cap \Sigma'''$ in (35) and the performance under sparse support sets given by $\Sigma = \Sigma''$ in (34). The purpose is to illustrate the advantage of sparse support sets with total support over those without total



Figure 6: Top: the point sets Y', Z' with n = 2500, respectively. Bottom: the point sets Y, Z with n = 5000, respectively. The color on Y, Y' illustrates the values $\nu^{(1)}$ and $\nu^{(1)'}$. The color on Z, Z' illustrates the values $\nu^{(2)}$ and $\nu^{(2)'}$.

n	Primal-dual	Gurobi-Barrier	SNNE	SNNE-t	SNNE-sparse	SNNE-sparse-2
250	3.6	19.1	17.3	1.3	2.2	1.0
500	40.6	77.3	134.3	4.1	6.1	4.4
800	141.3	236.2	380.7	11.8	14.3	11.3
1000	280.0	375.5	605.3	24.3	23.6	16.8
2500	5319	2834	4636	277.0	325.3	106.3
5000	44230	17180	18740	1136	1452	504.0
12500	> 50000	MemoryError	> 50000	13959	12001	6502

Table 5: Computational time(sec) based on rigid error reaching 10^{-4} .

Table 6: Matrix balancing error $||Mx - \mathbb{1}_{2n}||$.

$\xi_{\rm max}$	(i) $\Sigma, K = 20$	(ii) $\Sigma, K = 80$	(iii) $\Sigma'', K = 20$	$(iv)\Sigma'', K = 80$	(v) Π ₀
2	2.27×10^{-5}	2.81×10^{-6}	inf	2.55×10^2	1.94×10^{-5}
4	5.56×10^{-6}	3.98×10^{-6}	inf	\inf	1.55×10^{-5}

Table 7: Objective values $\mathbb{F}_t(Q, x)$. Here "NaN" stands for "Not a number".

$\xi_{\rm max}$	(i) $\Sigma, K = 20$	(ii) $\Sigma, K = 80$	(iii) $\Sigma'', K = 20$	(iv) $\Sigma'', K = 80$	(v) Π_0
2	5.81×10^5	1.57×10^6	NaN	$1.26 imes 10^6$	1.02×10^7
4	$6.35 imes 10^5$	$1.48 imes 10^6$	NaN	NaN	$1.02 imes 10^7$

support. As a reference, we also conduct the simulation with $\Sigma = \Pi_0$, i.e., the original complete index set as the support.

Consider the minimization in (122) with t = 1, n = 2500. Fix Q = I. Use the NE method to compute ν and update Σ according to the following five rules, including

- (i) Sparse support set Σ with K = 20 in (35);
- (ii) Sparse support set Σ with K = 80 in (35);
- (iii) Sparse support set $\Sigma = \Sigma''$ with K = 20 in (34);
- (iv) Sparse support set $\Sigma = \Sigma''$ with K = 80 in (34;
- (v) The complete set $\Pi_0 := \{(i, j) : i = 1, \dots, n, j = 1, \dots, n\}$.

Repeat the above (ν, Σ) -procedure ξ_{max} times to get an approximate minimizer x for (122). Results are reported in Table 6-8.

Table 6 reports the matrix balancing error of x. For (i),(ii) and (v), the support sets have total support and we can obtain accurate matrix balancing in these cases. Since t = 1 is used, the approximate solution x is far from a permutation solution and we are not concerned with accurate objective values. Hence, it is not surprisingly to see some numerical gap in Table 7, when objective values in (i),(ii) and (v) are compared. Indeed, as the size of support increases, more positive terms in $\langle \mathbb{T}^{-1}(\mathbb{1}_{\Sigma}), \log x \rangle$ contribute to the increase of objective values. Lastly, Table 8 reports the norm of the null vector $M^{\top}\nu$. In these three cases, the norm of the corresponding dual vectors are of similar size $\sim 10^3$.

On the other hand, since the set in (iii) or (iv) does not have total support, we can not get accurate matrix balancing to produce acceptable objective values. High accurate matrix balancing here is a very challenging task. Due to lack of total support, we also observe the blow-up of the dual vector norm. The norm is of size ~ 10⁵. See (iii) and (iv) in Table 8. Under this circumstance, the vector ν with very large norm can easily ruin the computational accuracy of the exponential functions in x.

$\xi_{ m max}$	(i) $\Sigma, K = 20$	(ii) $\Sigma, K = 80$	(iii) $\Sigma'', K = 20$	(iv) $\Sigma'', K = 80$	(v) Π_0
2	3.30×10^{3}	3.42×10^3	7.43×10^{5}	7.96×10^{5}	3.98×10^3
4	4.72×10^3	3.47×10^3	7.42×10^5	7.63×10^5	3.98×10^3

Table 8: The Frobenius norm $||M^{\top}\nu||_F$ of dual vectors.

4.3 Conclusion

Optimal transport, which is one assignment problem, can be handled by many methods, including the dual simplex method and the primal-dual methods. With negative entropy regularization, we can use matrix balancing algorithms to reach one approximate solution to optimal transport. In the study, we are concerned with Newton method based matrix balancing algorithms to point-set matching problems, i.e., SNNE and SNNE-sparse methods. One advantage of SNNE is that the method solely updates multiplier vectors along the increase of t, i.e., no need to store/pass x between each sub-problem. With the aid of sparse support, SNNE-sparse can be a relatively convenient tool in solving large-scale point-set matching problems. To ensure the solution quality from matrix balancing, we employ one simple rule to update these sparse support sets, in order to meet total support condition. With the aid of total support assumption, we can establish the convergence of LB and its step size analysis, which sheds light on the convergence of KR.

4.4 Data availability

The teapot dataset can be retrieved from the matlab 3-D point cloud file, "pcread('teapot.ply')". The lung branch points of subject H6012 is available from the corresponding author upon request.

4.5 Acknowledgements

We thank anonymous referees for helpful comments and suggestions that lead to improvement of the original manuscript.

A Appendix

A.1 Consistency of (20)

For x > 0, the null space of $M \operatorname{diag}(x)^2 M^{\top}$ has dimension 1.

Proposition A.1. Consider a positive vector $x \in \mathbb{R}^{n^2}$ and a matrix M in (4). Then Mdiag(x) has rank 2n-1 and

$$null(M \operatorname{diag}(x)^2 M^{+}) = null(M^{+}) = span\{[\mathbb{1}_n; -\mathbb{1}_n]\}.$$
(125)

In addition, for each $r \in \mathbb{R}^{n^2}$ and $x \in \mathbb{T}^{-1}(\Pi_n)$, the system

$$M \operatorname{diag}(x^2) M^{\dagger} u = M \operatorname{diag}(x^2) r \tag{126}$$

is consistent.

Proof. Suppose $M \operatorname{diag}(x)^2 M^{\top} u = 0$ for some $u \in \mathbb{R}^{2n}$. Then

$$0 = \langle u, M \operatorname{diag}(x)^2 M^{\mathsf{T}} u \rangle = \| \operatorname{diag}(x) M^{\mathsf{T}} u \|^2$$
(127)

implies diag $(x)M^{\top}u = 0$, i.e., $M^{\top}u = 0$. Hence, $null(M \text{diag}(x)^2 M^{\top}) \subseteq null(M^{\top})$. Besides, write u = [v; w] with some vectors $v \in \mathbb{R}^n$ and $w \in \mathbb{R}^n$. Since $M^{\top}u = 0 = \mathbb{1}_n w^{\top} + v\mathbb{1}_n^{\top} = 0$, then $u_i + w_j = 0$ for all $i, j = 1, \ldots, n$, i.e., $u_i = u_1 = -w_j$ for all i, j. This establishes

$$null(M \operatorname{diag}(x)^2 M^{\top}) \subseteq null(M^{\top}) \subseteq span\{[\mathbb{1}_n; -\mathbb{1}_n]\}.$$

On the other hand, consider a vector in the form $u = c[\mathbb{1}_n; -\mathbb{1}_n]$ with $c \in \mathbb{R}$. Then $M^{\top}u = c(\mathbb{1}_n\mathbb{1}_n^{\top} - \mathbb{1}_n\mathbb{1}_n^{\top}) = 0$ and $u \in null(M \operatorname{diag}(x)^2 M^{\top})$. This completes the proof of the first part. Finally, note that (126) is the associated normal equation to the least squares problem

$$\min_{u} \|\operatorname{diag}(x)M^{\top}u - \operatorname{diag}(x)r\|^{2}.$$
(128)

Hence, (126) is consistent.

A.2 Early termination

The following rounding procedure could quickly provide a KKT candidate point before the degeneracy of Schur complement matrices occurs. Suppose that one diagonal in $\mathbb{T}(x^{(t)})$ dominates other diagonals for some t. Then we have early termination of the interior point method, i.e., a permutation matrix can be identified as one optimal solution from $\mathbb{T}(x^{(t)})$. For simplicity, the following discussion does not involve support constraints.

Proposition A.2. Let $\gamma' \in (0,1)$ and $\gamma'' \in (1,\infty)$. Let $\hat{\nu} = [\hat{\nu}^{(1)}, \hat{\nu}^{(2)}]$. Let $(\hat{x}, \hat{\nu})$ be one approximate KKT point to (12) for some t > 0 with the entry wise bounds

$$\gamma' t^{-1} \le \widehat{x} \odot \widehat{s} \le \gamma'' t^{-1}, \ \widehat{s} = c - M^\top \widehat{\nu},$$
(129)

Let $X := \mathbb{T}(\hat{x})$. Suppose that for some permutation $\mathcal{J} : \{1, 2, \dots, n\} \to \{1, 2, \dots, n\}$,

$$X_{i,j} \le \frac{\gamma'}{\gamma''} X_{i,\mathcal{J}(i)} \text{ for all } j \neq \mathcal{J}(i),$$
(130)

Let $\nu := [\nu^{(1)}; \nu^{(2)}] \in \mathbb{R}^{2n}$ be given by

$$\nu^{(1)}(i) := c_{i,\mathcal{J}(i)} - \nu^{(2)}(\mathcal{J}(i)), \text{ where } \nu^{(2)} := \hat{\nu}^{(2)}.$$
(131)

Let \widetilde{X} be the permutation,

$$\widetilde{X}_{i,\mathcal{J}(i)} = 1 \text{ and } \widetilde{X}_{i,j} = 0, \ j \neq \mathcal{J}(i)$$
 (132)

Then (\tilde{x}, ν) is one KKT point to (7), where $\tilde{x} := \mathbb{T}^{-1}(\tilde{X})$.

Proof. The condition in (129) ensures that for all i, j = 1, ..., n,

$$\epsilon_{i,j} := t X_{i,j} (c_{i,j} - (M^\top \widehat{\nu})_{i,j}) \in (\gamma', \gamma'').$$
(133)

In particular, for $j = \mathcal{J}(i)$,

$$c_{i,\mathcal{J}(i)} - \hat{\nu}^{(1)}(i) - \hat{\nu}^{(2)}(\mathcal{J}(i)) \le \gamma''(tX_{i,\mathcal{J}(i)})^{-1}.$$
 (134)

We shall prove that (7) holds under this ν . Let $s := c - M^{\top} \nu$. From the definition in (131), it suffices to show $s_{i,j} \ge 0$ for all entries with $j \ne \mathcal{J}(i)$. From (133) and (131), we have

$$s_{i,j} := c_{i,j} - (M^{\top}\nu)_{i,j} = c_{i,j} - \nu^{(1)}(i) - \nu^{(2)}(j)$$
(135)

$$\geq c_{i,j} - c_{i,\mathcal{J}(i)} + \hat{\nu}^{(2)}(\mathcal{J}(i)) - \hat{\nu}^{(2)}(j) - \hat{\nu}^{(1)}(i) + \hat{\nu}^{(1)}(i)$$
(136)

$$\geq (tX_{i,j})^{-1} \epsilon_{i,j} - \gamma'' \frac{X_{i,j}}{X_{i,\mathcal{J}(i)}} (tX_{i,j})^{-1}$$
(137)

$$\geq (tX_{i,j})^{-1} (\epsilon_{i,j} - \gamma') \ge 0, \tag{138}$$

where we used the assumption in (130) and (134).

A.3 Proof of Theorem 3.

We shall prove Theorem 3. Recall B_k and C_k in (83,85). In addition to u_k in (88), introduce a few notations:

$$\lambda_k^2 := \langle (B_k - I_{2n}) \mathbb{1}_{2n}, (C_k + B_k)^{\dagger} (B_k - I_{2n}) \mathbb{1}_{2n} \rangle,$$
(139)

$$v_k := (B_k - I_{2n}) \mathbb{1}_{2n}, \ y_k := -(C_k + B_k)^{\dagger} (B_k - I_{2n}) \mathbb{1}_{2n}.$$
(140)

The LB iteration ζ_{k+1} is given by

$$\zeta_{k+1} = \zeta_k \odot + \alpha_k u_k = \zeta_k \odot (1 - \alpha_k y_k) \tag{141}$$

for some step size α_k within $(0, y_{\max}^{-1}(1 - \epsilon_+))$, where the safeguard parameter ϵ_+ and y_{\max} are defined in Remark 3.4. Before we proceed, we prove one crucial property: positive upper bounds exist for $\{\|C_k + B_k\|\}_{k=1}^{\infty}$ and $\{\|(C_k + B_k)^{\dagger}\|\}_{k=1}^{\infty}$. Let ζ_* be one minimizer of $\mathbf{g}(\zeta)$ and ζ_1 be one starting point of LB. We introduce a set of matrices,

$$S = \{ (\boldsymbol{\zeta}^{(1)} \boldsymbol{\zeta}^{(2)}^{\top}) \odot \mathbb{1}_{\Sigma} \in \mathbb{R}^{n \times n} : \mathbf{g}(\boldsymbol{\zeta}_1) \ge \mathbf{g}(\boldsymbol{\zeta}) \ge \mathbf{g}(\boldsymbol{\zeta}_*), \boldsymbol{\zeta} = [\boldsymbol{\zeta}^{(1)}; \boldsymbol{\zeta}^{(2)}] > 0 \}.$$
(142)

Then S is compact from Prop. 3.7. Introduce $H(\zeta)$ in (94). Note that $H(\zeta_k) = C_k + B_k$. Let $w = [w^{(1)}; w^{(2)}]$. We have norm estimates for $H(\zeta)$,

$$\|H(\zeta)\| \le \max_{w} \|w\|^{-2} \left(\sum_{j=1}^{n} \sum_{i=1}^{n} A_{i,j} \zeta_{i}^{(1)} \zeta_{j}^{(2)} (w_{i}^{(1)} - w_{j}^{(2)})^{2} \right)$$
(143)

$$\leq \max_{w} \|w\|^{-2} \left(\sum_{i=1}^{n} \sum_{j=1}^{n} A_{i,j} \zeta_{i}^{(1)} \zeta_{j}^{(2)} (1^{2} + 1^{2}) (w_{i}^{(1)^{2}} + w_{j}^{(2)^{2}}) \right)$$
(144)

$$= 2 \max\{\|(\operatorname{diag}(\zeta^{(1)}) A \operatorname{diag}(\zeta^{(2)}))\mathbb{1}_n\|_{\infty}, \|(\operatorname{diag}(\zeta^{(1)}) A \operatorname{diag}(\zeta^{(2)}))^{\top} \mathbb{1}_n\|_{\infty}\}.$$
 (145)

We have the following upper bound,

$$\|H(\zeta)\| \le 2\left(\max_{(i,j)\in\Sigma} A_{i,j}\right) \max\left(\|(\zeta^{(1)}\zeta^{(2)^{\top}} \odot \mathbb{1}_{\Sigma})\mathbb{1}_n\|_{\infty}, \|(\zeta^{(1)}\zeta^{(2)^{\top}} \odot \mathbb{1}_{\Sigma})^{\top}\mathbb{1}_n\|_{\infty}\right).$$
(146)

Thanks to Prop. 3.7, a constant **M** exists as a upper bound for $||H(\zeta)||$. On the other hand, for each w, we can express

$$\langle w, H(\zeta)w \rangle = \langle A \odot (\zeta^{(1)} {\zeta^{(2)}}^{\top} \odot \mathbb{1}_{\Sigma}), (w^{(1)} \mathbb{1}_{n}^{\top} - \mathbb{1}_{n} w^{(2)}^{\top}) \odot (w^{(1)} \mathbb{1}_{n}^{\top} - \mathbb{1}_{n} w^{(2)}^{\top}) \rangle$$
(147)

as one function defined on S. The null space of $H(\zeta)$ is \mathcal{N} for each $\zeta > 0$ from Prop. 3.5. Consider the following function to characterize the smallest positive eigenvalue of $H(\zeta)$,

$$\widehat{H}(\zeta) := \min_{w} \left\{ \frac{\langle w, H(\zeta)w \rangle}{\|w\|^2} : w \text{ is orthogonal to } \mathcal{N} \right\} > 0.$$
(148)

Since S is compact, then a positive constant **m** exists as a lower bound for the smallest positive eigenvalue of $H(\zeta)$. Hence, $||H(\zeta)^{\dagger}|| \leq \mathbf{m}^{-1}$. In summary, for all k, we have

$$||C_k + B_k|| \le \mathbf{M}, ||(C_k + B_k)^{\dagger}|| \le \mathbf{m}^{-1}.$$
 (149)

In addition, $||B_k|| \leq \mathbf{M}_1$ holds for some constant \mathbf{M}_1 .

The convergence of LB can be established by standard arguments in section 9.5 in [BV04]. Introduce a function of α ,

$$\widetilde{\mathbf{g}}(\alpha) := \mathbf{g}(\zeta_k + \alpha u_k). \tag{150}$$

Let $\tilde{\mathbf{g}}'$ and $\tilde{\mathbf{g}}''$ denote the first derivate and the second derivate of $\tilde{\mathbf{g}}$, respectively. Calculus shows

$$\nabla \mathbf{g}(\zeta) = \widetilde{A}\zeta - \zeta^{-1}, \ \nabla^2 \mathbf{g}(\zeta) = \widetilde{A} + \operatorname{diag}(\zeta^{-2}).$$
(151)

Proposition A.3 (Damped Newton phase). Let ϵ_+ be the safeguard parameter in Remark 3.4. Then

$$\lim_{k \to \infty} \lambda_k = 0, \ \lim_{k \to \infty} \|y_k\| = 0, \ \lim_{k \to \infty} \|v_k\| = 0.$$
(152)

Proof. First, we show that the limit of step size interval in Remark 3.4 is not zero. Indeed, since $y_{\max} \leq ||y_k|| \leq ||(C_k + B_k)^{\dagger}|| ||(B_k - I_{2n}) \mathbb{1}_{2n}|| \leq \mathbf{m}^{-1} (\mathbf{M}_1 + 1) ||\mathbb{1}_{2n}||$ for each ζ_k with $\mathbf{g}(\zeta_k) \leq c_0$, then $y_{\max}^{-1}(1 - \epsilon_+)$ stays away from 0 for each iteration. Second, we show

$$\widetilde{\mathbf{g}}(\alpha) - \widetilde{\mathbf{g}}(0) \le (-\alpha + \frac{\alpha^2}{2} (\mathbf{M}_1 + \epsilon_+^{-2}) \mathbf{m}^{-1}) \lambda_k^2.$$
(153)

Indeed, Taylor's formula indicates that for some scalar $\tilde{\alpha} \in [0, \alpha]$,

$$\widetilde{\mathbf{g}}(\alpha) = \widetilde{\mathbf{g}}(0) + \widetilde{\mathbf{g}}'(0)\alpha + \widetilde{\mathbf{g}}''(\widetilde{\alpha})\frac{\alpha^2}{2}$$
(154)

$$\leq \mathbf{g}(\zeta) + \alpha \nabla \mathbf{g}(\zeta)^{\top} u_k + \frac{\alpha^2}{2} \| \operatorname{diag}(\zeta_k) (\nabla^2 \mathbf{g}(\zeta_k + \widetilde{\alpha} u_k)) \operatorname{diag}(\zeta_k) \| \| y_k \|^2$$
(155)

$$\leq \mathbf{g}(\zeta) + \alpha(-\lambda_k^2) + \frac{\alpha^2}{2} (\|B_k\| + \|(1 + \widetilde{\alpha}y_k)^{-2}\|_{\infty}) \|y_k\|^2$$
(156)

$$\leq \widetilde{\mathbf{g}}(0) + \alpha(-\lambda_k^2) + \frac{\alpha^2}{2} (\mathbf{M}_1 + \epsilon_+^{-2}) \mathbf{m}^{-1} \lambda_k^2.$$
(157)

Together, the step size α_k in backtracking line search is bounded below by some positive constant. Since $\mathbf{g}(\zeta)$ is bounded below, then λ_k^2 must tend to 0, as $k \to \infty$. From (140), we have $\|v_k\|^2 \leq \lambda_k^2 \|C_k + B_k\| \leq \mathbf{M}\lambda_k^2$ and $\|y_k\|^2 \leq \|(C_k + B_k)^{\dagger}\|\lambda_k^2 \leq \mathbf{m}^{-1}\lambda_k^2$, which completes the proof. \Box

Proposition A.4 ($\alpha_k = 1$ phase). As k is sufficiently large, we have $\alpha_k = 1$.

Proof. Let $\epsilon_+ > 0$ be the safeguard parameter in Remark 3.4. Let $L = \epsilon_+^{-3}$. Since $u_k = -\zeta_k \odot y_k$,

$$|\widetilde{\mathbf{g}}''(\alpha) - \widetilde{\mathbf{g}}''(0)| \le |u_k^\top (\nabla^2 \mathbf{g}(\zeta_k + \alpha u_k) - \nabla^2 \mathbf{g}(\zeta_k))u_k|$$
(158)

$$\leq |u_k^{\top}(\operatorname{diag}(\zeta_k + \alpha u_k)^{-2} - \operatorname{diag}(\zeta_k^{-2}))u_k| = \left|y_k^{\top}\{(\frac{1}{1 - \alpha y_k})^2 - 1\}y_k\right| \leq \alpha L ||y_k||^3.$$
(159)

Hence, $\widetilde{\mathbf{g}}''(\alpha) \leq \widetilde{\mathbf{g}}''(0) + \alpha L \|y_k\|^3$. By integration, we have $\widetilde{\mathbf{g}}'(\alpha) \leq \widetilde{\mathbf{g}}'(0) + \alpha \widetilde{\mathbf{g}}''(0) + \frac{\alpha^2}{2} L \|y_k\|^3$, and

$$\widetilde{\mathbf{g}}(\alpha) - \widetilde{\mathbf{g}}(0) \leq \alpha \widetilde{\mathbf{g}}'(0) + \frac{\alpha^2}{2} \widetilde{\mathbf{g}}''(0) + \frac{\alpha^3}{6} L \|y_k\|^3$$
(160)

$$\leq -\alpha\lambda_{k}^{2} + \frac{\alpha^{2}}{2}(\lambda_{k}^{2} + \|v_{k}\|_{\infty}\|y_{k}\|^{2}) + \frac{\alpha^{3}}{6}L\|y_{k}\|^{3}$$
(161)

$$\leq \lambda^{2}(-\alpha + \frac{\alpha^{2}}{2}(1 + \|v_{k}\|_{\infty}\mathbf{m}^{-1}) + \frac{\alpha^{3}}{6}L\mathbf{m}^{-3/2}\lambda_{k})$$
(162)

where we used

$$\widetilde{\mathbf{g}}''(0) = \langle u_k, \nabla^2 \mathbf{g}(\zeta_k) u_k \rangle \tag{163}$$

$$= \langle (C_k + B_k)^{\dagger} (B_k - I_{2n}) \mathbb{1}_{2n}, (I_{2n} + B_k) (C_k + B_k)^{\dagger} (B_k - I_{2n}) \mathbb{1}_{2n} \rangle$$
(164)

$$= \lambda_k^2 + \langle (C_k + B_k)'(B_k - I_{2n})\mathbb{1}_{2n}, (I_{2n} - C_k)(C_k + B_k)'(B_k - I_{2n})\mathbb{1}_{2n} \rangle$$
(165)

$$\leq \lambda_k^2 + \|C_k - I_{2n}\| \|y_k\|^2 = \lambda_k^2 + \|B_k \mathbb{1}_{2n} - \mathbb{1}_{2n}\|_{\infty} \|y_k\|^2 = \lambda_k^2 + \|v_k\|_{\infty} \|y_k\|^2.$$
(166)

Take $\alpha = 1$ in (162). Using $||y_k||^2 \le \lambda_k^2 \mathbf{m}^{-1}$ and $||v_k||_{\infty} \le ||v_k|| \le \mathbf{M}^{1/2} \lambda_k$ from (149), we have

$$\widetilde{\mathbf{g}}(1) - \widetilde{\mathbf{g}}(0) \le \frac{\lambda_k^2}{2} \left(-1 + (\mathbf{M}^{1/2}\mathbf{m}^{-1} + \frac{\alpha^3}{3}L\mathbf{m}^{-3/2})\lambda_k \right)$$
(167)

Note that $\lim_{k\to\infty} \lambda_k = 0$ from Prop. A.3. When λ_k is sufficiently close to 0, $\alpha_k = 1$ is accepted by the backtracking line search. That is, for k sufficiently large, (167) indicates that

$$\widetilde{\mathbf{g}}(1) - \widetilde{\mathbf{g}}(0) \le \beta \nabla \mathbf{g}(\zeta_k)^\top u_k = -\beta \lambda_k^2$$
(168)

holds with backtracking parameter $\beta \in (0, 1/2)$.

- 1	-	-	٦
	-	_	_

References

[AZLOW17] Zeyuan Allen-Zhu, Yuanzhi Li, R. Oliveira, and A. Wigderson. Much faster algorithms for matrix scaling. 2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS), pages 890–901, 2017.

- [Bac70] M. Bacharach. *Biproportional Matrices and Input-Output Change*. Cambridge. University. Department of Applied Economics. 16 Monographs. Cambridge University Press, 1970.
- [BCC⁺15] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative bregman projections for regularized transportation problems. SIAM Journal on Scientific Computing, 37(2):A1111–A1138, 2015.
- [BCLW17] C. Brauer, Christian Clason, Dirk A. Lorenz, and Benedikt Wirth. A sinkhorn-newton method for entropic optimal transport. arXiv: Optimization and Control, 2017.
- [BDM09] Rainer Burkard, Mauro Dell'Amico, and Silvano Martello. Assignment Problems. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2009.
- [Ber03] D. P. Bertsekas. Nonlinear Programming. Athena Scientific, 2003.
- [BM92] P. J. Besl and N. D. McKay. A method for registration of 3-D shapes. IEEE Trans. Pattern Anal. Mach. Intell., 14:239–256, 1992.
- [BV04] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
- [CA14] Marco Cuturi and David Avis. Ground metric learning. J. Mach. Learn. Res., 15(1):533–564, jan 2014.
- [Che11a] P. Chen. A novel kernel correlation model with the correspondence estimation. *JMIV*, 39(2):100–120, 2011.
- [Che11b] P. Chen. A novel kernel correlation model with the correspondence estimation. *Journal of mathematical imaging and vision*, In press 2011.
- [CLC13] Pengwen Chen, Ching-Long Lin, and I-Liang Chern. A perfect match condition for point-set matching problems using the optimal mass transport approach. *SIAM Journal on Imaging Sciences*, 6(2):730–764, 2013.
- [CMTV17] Michael B. Cohen, A. Madry, D. Tsipras, and Adrian Vladu. Matrix scaling and balancing via box constrained newton's method and interior point methods. 2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS), pages 902–913, 2017.
- [CPSV18] Lenaic Chizat, Gabriel Peyre, Bernhard Schmitzer, and Francois-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Math. Comp.*, 87(314):2563–2609, 2018.
- [CR00] H. Chui and A. Rangarajan. A new algorithm for non-rigid point matching. *CVPR*, 2:44–51, 2000.
- [Cut13] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NIPS*, 2013.
- [Eva97] L. C. Evans. Partial Differential equations methods and Monge-Kantorovich mass transfer. ed. by S.T. Yau, International Press, Boston, 1997.
- [FM68] A. Fiacco and G. McCormick. Nonlinear programming;: Sequential unconstrained minimization techniques. 1968.
- [FO08] Haw-ren Fang and Dianne P. O'Leary. Modified cholesky algorithms: a catalog with new approaches. *Mathematical Programming*, 115(2):319–349, 2008.
- [Gon12] Jacek Gondzio. Interior point methods 25 years later. European Journal of Operational Research, 218(3):587–601, 2012.

- [GTY04] J. Glaunes, A. Trouve, and L. Younes. Diffeomorphic matching of distributions: A new approach for unlabelled point-sets and sub-manifolds matching. *CVPR*, 2:712–718, 2004.
- [GWXY19] Dongdong Ge, Haoyue Wang, Zikai Xiong, and Yinyu Ye. Interior-Point Methods Strike Back: Solving the Wasserstein Barycenter Problem. Curran Associates Inc., Red Hook, NY, USA, 2019.
- [Ide16] Martin Idel. A review of matrix scaling and sinkhorn's normal form for matrices and positive maps. *arXiv: Rings and Algebras*, 2016.
- [Kai98] T. Kaijser. Computing the Kantorovich distance for images. J. Math. Imaging and Vision, 9:173–191, 1998.
- [Kan42] L. V. Kantorovich. On the transfer of masses. Dokl. Akad. Nauk. SSSR, 37:227–229, 1942.
- [KK92] Leonid Khachiyan and Bahman Kalantari. Diagonal matrix scaling and linear programming. SIAM Journal on Optimization, 2(4):668–672, 1992.
- [KLRS08] B. Kalantari, I. Lari, F. Ricca, and B. Simeone. On the complexity of general matrix scaling and entropy minimization via the ras algorithm. *Mathematical Programming*, 112(2):371– 401, 2008.
- [Kni08] Philip A. Knight. The Sinkhorn–Knopp algorithm: Convergence and applications. SIAM Journal on Matrix Analysis and Applications, 30(1):261–275, 2008.
- [KPT⁺17] Soheil Kolouri, Se Rim Park, Matthew Thorpe, Dejan Slepcev, and Gustavo K. Rohde. Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine*, 34(4):43–59, 2017.
- [KR12] Philip A. Knight and Daniel Ruiz. A fast algorithm for matrix balancing. IMA Journal of Numerical Analysis, 33(3):1029–1047, 10 2012.
- [KR17] Johan Karlsson and Axel Ringh. Generalized sinkhorn iterations for regularizing inverse problems using optimal mass transport. *SIAM J. Imaging Sci.*, 10(4):1935–1962, 2017.
- [KS67] Paul Knopp and Richard Sinkhorn. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343 348, 1967.
- [LY16] David G. Luenberger and Yinyu Ye. *Linear and Nonlinear Programming*. Springer International Publishing, 2016.
- [Meh92] Sanjay Mehrotra. On the implementation of a primal-dual interior point method. SIAM Journal on Optimization, 2(4):575–601, 1992.
- [MO68] Albert W. Marshall and Ingram Olkin. Scaling of matrices to achieve specified row and column sums. *Numerische Mathematik*, 12(1):83–90, 1968.
- [MSKL09] O. Museyko, M. Stiglmayr, K. Klamroth, and G. Leugering. On the application of the Monge-Kantorovich problem to image registration. SIAM J. Imaging Sciences, 2(4):1068– 1097, 2009.
- [MV98] J.B.Antoine Maintz and Max A. Viergever. A survey of medical image registration. *Medical Image Analysis*, 2(1):1 36, 1998.
- [PC19] Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. Foundations and Trends® in Machine Learning, 11(5-6):355–607, 2019.
- [RDG09] J. Rabin, J. Delon, and Y. Gousseau. A statistical approach to the matching of local features. SIAM J. Imaging sciences, 2:931–958, 2009.
- [Rob12] Robert Robere. Interior point methods and linear programming. 2012.

- [RTG00a] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. Int. J. Comput. Vis., 40:99–121, 2000.
- [RTG00b] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.
- [Sch19] Bernhard Schmitzer. Stabilized sparse scaling algorithms for entropy regularized transport problems. *SIAM Journal on Scientific Computing*, 41(3):A1443–A1481, 2019.
- [Sin64] R. Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. Ann. Math. Statist., 35:876–879, 1964.
- [SS15] Bernhard Schmitzer and Christoph Schnörr. Globally optimal joint image segmentation and shape matching based on wasserstein modes. *Journal of Mathematical Imaging and Vision*, 52(3):436–458, 2015.
- [Vil03] C. Villani. *Topics in Optimal Transportation*. Graduate Studies in Mathematics, AMS, 2003.
- [Vil08] C. Villani. *Optimal transport: Old and New.* Springer Verlag (Grundlehren der mathematischen Wissenschaften), 2008.
- [Wah90] G. Wahba. Spline models for observational data. SIAM, Philadelphia, PA, 1990.
- [WPR85] M. Werman, S. Peleg, and A. Rosenfeld. A distance metric for multi-dimensional histograms. Comp. Vis. Graphics Image Proc., 32:328–336, 1985.
- [Wri97] Stephen J. Wright. *Primal-dual interior-point methods*. Society for Industrial and Applied Mathematics, 1997.
- [YLST21] Lei Yang, Jia Li, Defeng Sun, and Kim-Chuan Toh. A fast globally linearly convergent algorithm for the computation of wasserstein barycenters. J. Mach. Learn. Res., 22:21:1–21:37, 2021.
- [ZF03] B. Zitova and J. Flusser. Image registration methods: a survey. *Image and Vis. Compu.*, 21:977–1000, 2003.
- [ZYHT07] L. Zhu, Y. Yang, S. Haker, and A. Tannenbaum. An image morphing technique based on optimal mass preserving mapping. *IEEE Image Processing*, 16(6):1481 1495, 2007.